

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

**A Thesis Submitted for the Degree of PhD at the University of Warwick**

<http://go.warwick.ac.uk/wrap/70995>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.



# Modelling Shape Fluctuations During Cell Migration

by

**Samuel David Russell Jefferyes**

**Thesis**

Submitted to the University of Warwick

for the degree of

**Doctor of Philosophy**

**Systems Biology**

September 2014

THE UNIVERSITY OF  
**WARWICK**

# Contents

<b>List of Tables</b>	<b>iv</b>
<b>List of Figures</b>	<b>v</b>
<b>Acknowledgments</b>	<b>vii</b>
<b>Declarations</b>	<b>viii</b>
<b>Abstract</b>	<b>ix</b>
<b>Abbreviations</b>	<b>x</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
1.1 Cell Migration . . . . .	1
1.1.1 Background to Cell Migration . . . . .	1
1.1.2 Hierarchical Reductionism . . . . .	2
1.1.3 Using Morphology to Study Migration . . . . .	3
1.2 Cell Shape Modelling . . . . .	3
1.2.1 Qualitative Shape Analysis in Literature . . . . .	3
1.2.2 Quantitative Shape Modelling in Literature . . . . .	4
1.3 Retinal Pigment Epithelial Cells . . . . .	6
1.4 Our Approach . . . . .	7
1.5 Thesis Organisation . . . . .	10
<b>Chapter 2 Materials and Methods</b>	<b>11</b>
2.1 Cell Culture and Imaging . . . . .	11
2.1.1 Cell Culture . . . . .	11
2.1.2 Imaging . . . . .	11
2.2 Cell Segmentation . . . . .	12
2.3 Diffusion Maps . . . . .	12

2.3.1	Laplace-Beltrami Normalisation . . . . .	13
2.4	Square-Root Elastic (SRE) distance . . . . .	13
2.5	Affinity Propagation . . . . .	14
2.6	Seriation Algorithm . . . . .	15
2.7	Hidden Markov Models . . . . .	16
2.8	Cell Track Data . . . . .	17
<b>Chapter 3</b>	<b>The Best Alignment Metric</b>	<b>18</b>
3.1	Introduction . . . . .	18
3.2	Shape Difference Metric . . . . .	18
3.2.1	Understanding the data . . . . .	18
3.2.2	Introducing the Best Alignment Metric . . . . .	20
3.2.3	Discussion of the Best Alignment Metric . . . . .	22
3.2.4	BAM comparisons . . . . .	23
3.3	Kernel Bandwidth . . . . .	26
3.4	Examining the Performance of the Best Alignment Metric . . . . .	28
3.4.1	Affinity Propagation for Independent Validation . . . . .	28
3.4.2	SRE versus BAM . . . . .	28
3.4.3	Affinity Propagation on a Large Dataset . . . . .	33
3.4.4	Seriation Extension to Affinity Propagation . . . . .	33
3.5	Application to Breast Cancer Histology Images . . . . .	35
3.5.1	Introduction . . . . .	35
3.5.2	Applying the Extended Affinity Propagation to Histology Data . . . . .	37
3.6	Discussion . . . . .	39
<b>Chapter 4</b>	<b>Morphological Phenotyping of Retinal Pigment Epithelial Cells</b>	<b>42</b>
4.1	Visualising Shape Space . . . . .	42
4.1.1	Shape Averaging . . . . .	42
4.1.2	Extended Affinity Propagation to Visualise DM Embedding . . . . .	43
4.2	Shape Feature Correlation . . . . .	46
4.2.1	Scalar Shape Features . . . . .	46
4.2.2	Shape feature correlation . . . . .	48
4.2.3	Features distributed over the Diffusion Maps embedding . . . . .	51
<b>Chapter 5</b>	<b>Turn Prediction</b>	<b>69</b>
5.1	Introduction . . . . .	69
5.1.1	Proposed Pipeline . . . . .	70



5.2	Development . . . . .	71
5.2.1	Morphological Analysis . . . . .	71
5.2.2	Migrational Analysis . . . . .	75
5.2.3	Turn Prediction Accuracy . . . . .	76
5.3	Results . . . . .	77
5.4	Discussion . . . . .	81
<b>Chapter 6</b>	<b>Discussion and Conclusions</b>	<b>84</b>
6.1	General Discussion . . . . .	84
6.2	Shape Analysis . . . . .	85
6.2.1	Shape Space Learning with Diffusion Maps . . . . .	85
6.2.2	Best Alignment Metric . . . . .	87
6.3	Mechanisms of migration . . . . .	88
6.3.1	Turn Prediction . . . . .	88
6.3.2	Our Migration Analysis in Context . . . . .	88
6.4	Plan for Publication . . . . .	90
<b>Appendix A</b>	<b>Best Alignment Metric</b>	<b>91</b>
A.1	Best Alignment Metric Formulation . . . . .	91
A.1.1	Dealing with $\phi$ . . . . .	92
A.1.2	Dealing with $r$ . . . . .	93
A.2	Planar translation to minimise pairwise distances between two sets . . . . .	95

# List of Tables

3.1	A table displaying the average time taken to compute SRE and BAM measurements of pairs of shapes, for a range of values for $N$ , which is the number of points sampled from around each curve. Each result is the average measurement computed over 5 distinct pairs of shapes.	30
4.1	This table shows the correlation of simple shape features with the Diffusion Maps representation of our RPE1 dataset. The shape features are described in section 4.2.1. The Diffusion embedding is described in section 2.3 . . . . .	48

# List of Figures

1.1	Example migrating epithelial cells. . . . .	7
1.2	Algorithm flow diagram. . . . .	9
3.1	Cell shape misalignment. . . . .	19
3.2	Comparative performance of shape distance measures. . . . .	25
3.3	Absolute relative error from curve sub-sampling. . . . .	29
3.4	Metric Comparison. . . . .	32
3.5	Affinity Propagation Exemplars . . . . .	34
3.6	Seriation Ordered Exemplars . . . . .	36
3.7	Example Mitotic Cell Candidates. . . . .	37
3.8	Ordered Exemplars for Mitosis Dataset. . . . .	38
3.9	Relative Mitotic Count in Ordered Clusters. . . . .	40
4.1	Axis phenotypes. . . . .	44
4.2	Affinity Propagation Clusters over DM Embedding. . . . .	45
4.3	Area distributed over embedding. . . . .	51
4.4	Major Axis Length distributed over embedding. . . . .	52
4.5	Minor Axis Length distributed over embedding. . . . .	53
4.6	Eccentricity distributed over embedding. . . . .	54
4.7	Orientation distributed over embedding. . . . .	55
4.8	Convex Area distributed over embedding. . . . .	56
4.9	Solidity distributed over embedding. . . . .	57
4.10	Extent distributed over embedding. . . . .	58
4.11	Perimeter distributed over embedding. . . . .	59
4.12	Circularity distributed over embedding. . . . .	60
4.13	Symmetry distributed over embedding. . . . .	61
4.14	Max distance from centre distributed over embedding. . . . .	62
4.15	Min distance from centre distributed over embedding. . . . .	63
4.16	Min/max centre distance ratio distributed over embedding. . . . .	64

4.17	Irregularity distributed over embedding. . . . .	65
4.18	Irregularity2 distributed over embedding. . . . .	66
5.1	Track analysis pipeline. . . . .	71
5.2	Shape representation for Detecting Repolarisation Events. . . . .	73
5.3	Example cell track labelled for training. . . . .	74
5.4	Angle check difficulties. . . . .	76
5.5	Angle distributions. . . . .	78
5.6	Example repolarising tracks. . . . .	79
5.7	Example alternative tracks. . . . .	80
5.8	Turn prediction accuracy. . . . .	81

# Acknowledgments

I offer my sincere thanks to both of my supervisors; Dr Anne Straube and Dr Nasir Rajpoot. Both have continuously offered me invaluable guidance, support and critique. I thank Anne specifically for being ever insightful and enthusiastic in a project somewhat outside of her usual field, and for being tolerant of my bizarre working routines. I'd like to thank Nasir, for his continued insightful guidance even from afar.

I'm very grateful to Prof David Epstein for his considerable input and useful advice and discussion.

I'd like to thank my advisory committee Prof. Andrew McAinsh, Dr Till Bretschneider and Prof. Matthew Turner, for their time and guidance.

I greatly appreciate the many individuals in the System Biology DTC, BCB and COMBI groups who sat through my talks, showed interest in my work and offered valuable input.

Many thanks to the BBSRC for funding this project.

Finally, I thank my family and friends, who have been a constant source of support throughout my PhD.

# Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

The work presented (including data generated and data analysis) was carried out by the author.

Parts of this thesis have been published by the author:

Samuel D R Jefferyes, David B A Epstein, Anne Straube, and Nasir M Rajpoot. *A novel framework for exploratory analysis of highly variable morphology of migrating epithelial cells. Conference proceedings : ...* Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, 2013:34636, January 2013.

# Abstract

Cell migration is of crucial importance for many physiological processes, including embryonic development, wound healing and immune response. Defects in cell migration are the cause of chronic inflammatory diseases, mental retardation and cancer metastasis. Cell movement is driven by actin-mediated cell protrusion, substrate adhesion and contraction of the cell body.

The emergent behaviour of the intracellular processes described above is a change in the morphology of the cell. This inspires the main hypothesis of this work which is that there is a measurable relationship between cell morphology dynamics and migratory behaviour, and that quantitative models of this relationship can create useful tools for investigating the mechanisms by which a cell regulates its own motility.

Here we analyse cell shapes of migrating human retinal pigment epithelial cells with the aim to map cell shape changes to cellular behaviour. We develop a non-linear model for learning the intrinsic low-dimensional structure of cell shape space and use the resultant shape representation to analyse quantitative relationships between shape and migration behaviour. The biggest algorithmic challenge overcome in this thesis was developing a method for efficiently and appropriately measuring the shape difference between pairs of cells that may have come from independent image scenes. This difference measure must be capable of coping with the widely varying morphologies exhibited by migrating epithelial cells. We present a new, rapid, landmark-free, shape difference measure called the Best Alignment Metric (BAM). We show that BAM performs highly within our framework, generating a shape space representation of a very large dataset without any prior information on the importance of any given shape feature.

We demonstrate quantitative evidence for a model of cell turning based on repolarisation and discuss the impact our proposed framework could have on the continued study of migratory mechanisms.

# Abbreviations

AP	Affinity Propagation
BAM	Best Alignment Metric
DM	Diffusion Maps
DMEM	Dulbecco's Modified Eagle Medium
GFP	Green Fluorescent Protein
GTP	Guanine-5'-triphosphate
HMM	Hidden Markov Model
hTERT	Human telomerase reverse transcriptase
$L^2$ -norm	Lebesgue 2-norm
NLDR	Non-linear Dimensionality Reduction
PCA	Principal Component Analysis
$\mathbb{R}$	The space of real numbers
RNA	Ribonucleic Acid
RPE	Retinal Pigment Epithelial
RPE1 GLA6	Retinal Pigment Epithelial cells with mGFP-LifeAct
RSKNN	Reverse Soft K-Nearest Neighbour Density Estimation
SRE	Square-Root Elastic (distance)
SRV	Square-Root Velocity (framework)



# Chapter 1

## Introduction

### 1.1 Cell Migration

#### 1.1.1 Background to Cell Migration

Cell migration is of fundamental importance for embryonic development, immune response and wound healing [Keller, 2005; Theveneau and Mayor, 2011; Tarbashevich and Raz, 2010; Marelli-Berg et al., 2010; Abreu-Blanco et al., 2012]. Equally importantly, defective cell migration is a primary cause of disease: it enables tissue invasion and metastasis by cancer cells, chronic inflammation and mental retardation [Hanahan et al., 2000; Ridley et al., 2003; Roussos et al., 2011]. Cell migration is achieved through dynamic control of the cytoskeleton and regulated through the integration of many complex signalling pathways [Ridley et al., 2003]. The most common mode of cell migration is a crawling process that can be conceptualised as a cyclic process with four steps [Mitchison and Cramer, 1996; Horwitz and Webb, 2003; Lauffenburger and Horwitz, 1996]. Firstly, the front edge of the cell protrudes forward, secondly adhesions are formed that anchor the cell membrane to the substrate or neighbouring cells. Thirdly, the adhesions in the rear of the cell disassemble and finally, contraction of the cytoplasm results in the forward translocation of the cell body. The mechanical power for the crawling process comes from a treadmilling process of the actin network [Pollard and Borisy, 2003], where filaments polymerise and branch at the leading edge and depolymerise at the rear of the network and from myosin-mediated actin filament sliding that cause contractions.

Blebbing is another way that eukaryotic cells can move. This happens when hydrostatic forces in a local region of the cell are high enough to cause the membrane of the cell to detach from the cytoskeleton and to bulge outwards [Charras et al., 2005]. The subsequent reparations to the ruptured cytoskeleton (specifically the

actin cortex) can either stabilise or retract the bulbous protrusion [Charras and Paluch, 2008]. Global regulation of these local decisions results in net movement of the cell. The hydrostatic forces are largely generated by myosin-mediated contractility [Charras et al., 2005].

While the underlying motile mechanisms are believed to be well preserved across differing cell types, the outward dynamic behaviour can vary greatly. *Dictyostelium discoideum* is a species of unicellular organism able to perform both crawling and blebbing and known to adapt its motile behaviour to the environment they find themselves in [Charras and Paluch, 2008]. Fish keratocytes, which have a constant canoe-like shape [Goodrich, 1924], have a large actin network across the leading region (lamellipodium) that barely changes configuration as it migrates [Mogilner and Edelstein-Keshet, 2002]. This constancy means that fish keratocytes are some of the fastest and most directionally persistent moving cells [Keren et al., 2009]. Other cells, such as fibroblasts and neuronal growth cones, more dynamically turn protrusions on and off in localised regions to allow more adaptive and sensitive motility, but at the cost of speed and persistence. These differences are reflected in the morphology of the cells, fish keratocytes have one large lamellipodium that changes very little while fibroblasts exhibit highly fluctuating morphologies, stochastically protruding smaller lamellipodia and filopodia in response to their surroundings [Abercrombie et al., 1970].

The ability for a cell to change between different styles of motile behaviour is very important in a number of physiological processes. Transitions between non-motile and motile states such as in Epithelial-Mesenchymal transition [Yang and Weinberg, 2008] have been linked to embryogenesis, wound healing and cancer metastasis. The ability for a cell to transition between mechanisms of migration has been linked to cancer cell metastasis; it is believed that a metastatic cancer cell will need to pass through different environments and may need to use different migratory mechanisms for each [Wang et al., 2005; Friedl and Wolf, 2003].

### 1.1.2 Hierarchical Reductionism

A recent review of cytoskeletal models discusses the difficulties faced when exploring a physical understanding of cellular mechanics [Huber et al., 2013]. It describes how the challenge for physicists is the range of physical scales over which the components of the cytoskeleton interact. Creating a model which explicitly links monomers to polymers to networks to cells to tissue is an intractable task. The proposed solution is a scheme of hierarchical reductionism. This is the process of examining each level of complexity only in terms of the level below it (or perhaps the nearest two), in

order to build a full model inductively, while each individual model remains manageable. An example of this may be seen from an investigation of cell protrusions in terms of net growth rate of local actin networks. The actin network growth rate is an emergent property of the polymerisation process of individual actin filaments. It would be a much more complex task to model a protrusion in terms of the individual filaments, considering that their orientations and polymerisation rates are not independent. In turn, one can look at the way that cell morphology, a cellular phenomenon, governs its motile behaviour, which is how the cell explores its surroundings. It is this scopic level that we shall be focussing on in this thesis.

### 1.1.3 Using Morphology to Study Migration

Much work is being carried out to study the dynamics of various cytoskeletal components at the molecular level. There are indeed mathematical models for each step of cell migration: actin-mediated protrusion, adhesion, contraction. However an integrated model that describes cell migration as a whole requires much more knowledge about how these subprocesses integrate and how this integration is affected by environmental cues [Danuser et al., 2013]. We hypothesise that morphology can efficiently represent the emergent behaviour of the intracellular system, in the sense that, in order to induce migration, the intracellular mechanisms must cause a structural change. We explore quantitative models for aspects of morphology and motility, and ultimately the dependence between the two. With knowledge of the common morphological behaviour patterns that accompany any cellular event, we believe it will be possible to infer information about the internal structural dynamics involved with that event.

## 1.2 Cell Shape Modelling

### 1.2.1 Qualitative Shape Analysis in Literature

The earliest papers on fish keratocytes [Goodrich, 1924] observed their distinctive shapes and described them qualitatively as canoe-like with a large fan section. The authors even commented on the relation between the geometric orientation of the fan and the direction of motion, although this relationship was not quantitatively modelled until later [Lee et al., 1993; Keren et al., 2008].

In many cases in literature, morphological change is observed with little substantiation. Cooper and Schliwa [Cooper and Schliwa, 1986] observed that applying a current to fish epidermal cells in culture causes them to flock to the cathode. It was

claimed that the cells have unchanged morphology in this process, to substantiate this claim the reader is encouraged to assess the figures subjectively. Wang et al. [Wang et al., 2003] claimed that cell shape (as well as polarity), in human embryonic kidney tumour cells, is regulated by RhoGTPase-dependent regulation of the actin cytoskeleton. This claim was substantiated by the qualitative judgement of the presence or lack of protrusions or lamellipodia-like structures.

### 1.2.2 Quantitative Shape Modelling in Literature

Here, as with all areas of science, a quantitative understanding has many advantages. A quantitative model of shape and shape dynamics allows for objectivity and statistical validation. It also confers the ability to make graded commentary, in other words we become able to quantify a claim such as shape A is more irregular than shape B. One does not naturally think of *shape* as a quantitative concept, however there are many ways to create such a representation and the rest of this section will look at different methods that have been used to achieve this in literature.

#### Shape Features

Simple shape features include basic scalar properties of shape, such as length, area, perimeter, concavity, circularity and symmetry to name just a few<sup>1</sup>. The simplest way to create a quantitative model for shape analysis of a given biological object is to measure one of these simple shape features. This technique has been applied to many varying biological settings. For example, cell length has been found to be an informative feature in the life and function of *S. pombe* [Martin and Berthelot-Grosjean, 2009; Moseley et al., 2009]. Measures of symmetry in breast lesions have shown the ability to distinguish between benign and malignant tumours [Yang et al., 2009; Liney et al., 2006].

Simple features can also be brought into dynamic models. For example, in human epithelial cells, although maximum cell tail length remains unchanged, the lifetime of individual tails decreases dramatically after epigenetic treatment [Theisen et al., 2012].

It is possible to select the examined features to be precisely relevant to the investigation. Rangayyan and Nguyen [Rangayyan and Nguyen, 2007] focus on measures of self similarity for categorising contours of breast masses to assist in breast cancer diagnosis; they find that the combination of fractal dimension and fractional concavity yields the best results.

---

<sup>1</sup>While these features all have some degree of obvious intuitive definition, an explicit formulation would always be given in an application.

Tweedy et al. examine the Fourier power spectrum of curves, which is a set of features that look at the various levels of periodicity that exist in each curve, and the authors show successful use of these features to discover the modes of shape variation in chemotaxing *D. discoideum* [Tweedy et al., 2013]. Note that, although this analysis is based on feature descriptors and not explicit descriptors (explained later), the representation carries a lot of information that is not readily accessible. This is tackled through use of machine learning, which is a common approach for handling high-dimensional and other difficult data, described below.

## Machine Learning

The concept of *shape* has many different aspects and hence it can be considered as high dimensional. The difficulty is that any quantitative modelling of patterns of shape behaviour will run into difficulty if the shape representation has too many dimensions. This is because of what is known as *the curse of dimensionality*, which refers to the fact that many analytical difficulties scale exponentially as the dimensionality of the data increases. In this case it is a problem of sampling; a high dimensional space needs a far higher sample size to adequately populate the data space. So it is necessary to perform dimensionality reduction and one field that allows this is Machine Learning.

Machine learning, much like organic learning, involves exposing the ‘learner’ to a large number of examples of the objects of interest, in such a way that the ‘learner’ can develop an ‘understanding’ of the allowed variation within the example set. When used for dimensionality reduction, the algorithm will simplify the input by mapping it into low-dimensional space, attempting to maximise the amount of information preserved in each consecutive new dimension. Often the data is represented to high accuracy with a small number of new dimensions. Tweedy *et al.* [Tweedy et al., 2013] make use of Principal Component Analysis (PCA) [Pearson, 1901], a well known linear dimensionality reduction technique, to convert their 64-dimensional Fourier descriptors into 3 modes of variation, which account for over 90% of the shape variability.

PCA is a tried and tested technique, however it can only be relied upon if the high-dimensional structure is (at least approximately) linear. However, PCA will fail to describe the data if the high-dimensional geometry of the dataset is more complicated than linear. Bakal *et al.* [Bakal et al., 2007] use neural networks to determine the important aspects of cell morphology as controlled by certain gene networks in *Drosophila*. A neural network model is a machine learning technique inspired by a model for the learning mechanisms of the brain, which has layers of ‘neurons’ which

activate each other and pick up on regular patterns. They use neural networks in a *bag of features* approach, whereby they measure a large number of features (in many cases with overlapping information) and simply let the machine learning algorithm find the structure therein. They use 145 morphological features and 249 genetic treatment conditions, and then search for joint clusters as a *data-mining* approach to finding gene networks that control morphological change.

### Explicit Shape Descriptors

Measuring a specified feature or property can be very useful in answering specific questions about those features. Given a hypothesis or some *a priori* information about the involvement of a specific feature within a larger system, then measurement of that feature becomes important. However, with a more open-ended enquiry it is possible that leading the analysis with specific feature based analysis may cause one to miss information about the intrinsic underlying mechanisms.

In other imaging tasks, such as image retrieval and object classification, feature-based measurements have other problems relating to the fact that objects can have similar features whilst being visually very different.

This motivates the use of explicit shape descriptors. An explicit shape descriptor is any shape representation that is reversible, i.e. the original shape data is recoverable. The requirement for a representation to be reversible guarantees that the representation contains all of the information about the object it is trying to represent. For the most part, explicit shape descriptors will be very high dimensional, which means dimensionality reduction is necessary. Unless there is empirical evidence or a theoretical justification that the structure of the data space is linear, PCA will not be sufficient to characterise the geometry. Sparks and Madabhushi [Sparks and Madabhushi, 2013] use a non-linear dimensionality reduction (NLDR) technique, called Graph Embedding, to create an explicit shape descriptor to represent prostate gland lumina, whose shape is used by pathologists to grade prostate tumour malignancy. The authors demonstrate that classification in their shape representation space is effective in this grading task.

## 1.3 Retinal Pigment Epithelial Cells

We choose human retinal pigment epithelial (RPE) cells as a model system as they show a high variability in shape while undergoing extensive directionally persistent migration, showing all characteristics of primary human cells while being amenable

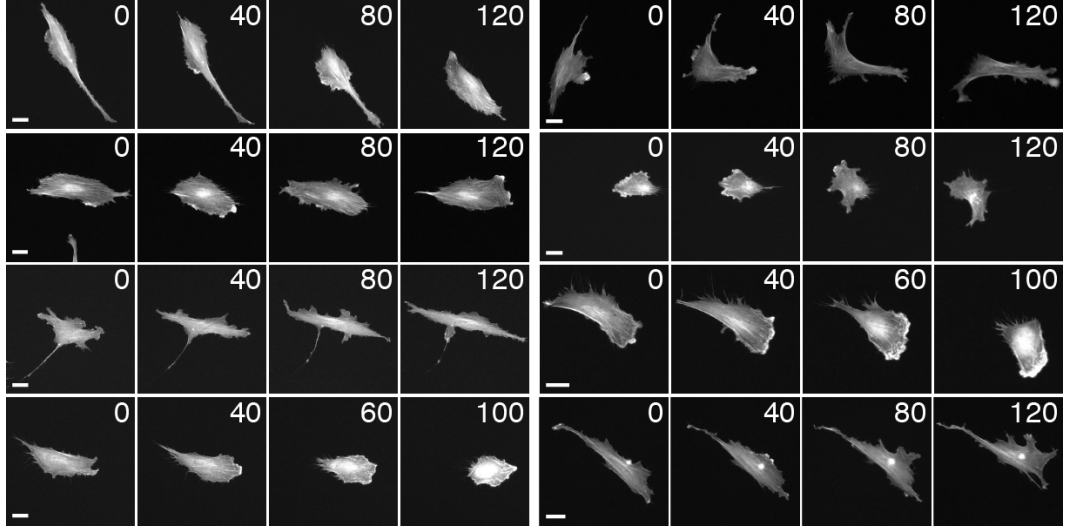


Figure 1.1: **Example migrating epithelial cells.** Representative frames of migrating RPE cells expressing mGFP-LifeAct to mark actin. Scale bars are  $20\mu\text{m}$  and relative time is indicated in minutes. Figure reproduced from [Jefferyes et al., 2013].

to genetic modification.

The retinal pigment epithelium is a monolayer that separates the photosensitive retina from the choroid in the eye. These cells perform many functions to maintain the visual performance of the photoreceptors [Strauss, 2005]. Since the epithelium exists as a monolayer *in vivo*, the cells readily take to a uniform 2D substrate *in vitro*. Importantly, RPE cells move freely in culture, without requiring stimulus to turn. This internally motivated behaviour is in contrast to the stimulated behaviour seen in chemoattraction experiments.

A high variability in cell shape is commonly found in images of cancer cell migration *in vivo* [Friedl and Wolf, 2003]. Although we simplify the system by looking at a 2D model, the complexity displayed in our dataset required the development of a novel shape comparison algorithm, tools that will be undoubtedly useful for tackling a three-dimensional model in the future.

## 1.4 Our Approach

The main research aim of the work presented in this thesis is to develop a framework for generating a quantitative representation of shape that is useful for modelling migrational behaviour in epithelial cells. A relatively straightforward way to achieve

that aim would be to use shape features that are known to be related to cell migration. However, we want to do this in such a way as to assume no prior knowledge of the importance of any individual shape features. This choice has several benefits. Firstly, it is immediately transferrable to other systems, and can provide information about which are the important features in systems where they are not known. Secondly, we do not limit our analysis to features that are known to be relevant, as it may be that other features are similarly important and their inclusion yields a richer analysis. Instead, we opt for a machine learning approach, which, as discussed earlier, generates a new description of the data that represents the prevalent modes of variation (or degrees of freedom) within the dataset. This approach will feed a descriptive model for the data that will present the shape distribution of our cells in a way that can be explored both subjectively, through visualisation techniques (see Chapter 4), and objectively as a quantitative representation of the cells that can be used to feed further models of dynamic behaviour (see Chapter 5).

Our framework follows on from work done by Rajpoot and Arif [Rajpoot and Arif, 2008], who use unsupervised learning to map the shape space of simple image outlines and successfully distinguish shapes such as guitars, apples, teddies, cars and carriages. They make use of the Diffusion Maps technique [Coifman and Lafon, 2006a], which seeks to learn the local structure of the data but attempts to ignore larger-scale geometry in the space of the descriptor. A crucial component in their framework is the shape similarity measure, since this measure of similarity is preserved in the final representation. A major task for this project was developing a similarity measure that appropriately and efficiently captures shape information and ignores extraneous information present in the image. Chapter 3 describes our work for this task, we discuss the requirements we perceived to be in place then show our own developments to solve this problem.

Figure 1.2 gives an overview of the generalised framework that is presented in this thesis. It shows that the cell contours are first segmented from the images, then converted into a descriptor representation. The dataset is then assimilated into a large shape similarity matrix, and from this matrix a low-dimensional, quantitative representation of the dataset is created.

We published this generalised framework, in the context of migrating human retinal pigment epithelial (RPE) cells, at the IEEE Engineering in Medicine and Biology Society conference in July 2013 [Jefferyes et al., 2013] using the Square-Root Velocity shape metric [Srivastava et al., 2011]. We have since then developed a novel shape comparison algorithm we term BAM: the Best Alignment Metric. This uses circular convolution to rapidly compute pairwise alignments and solve issues of in-



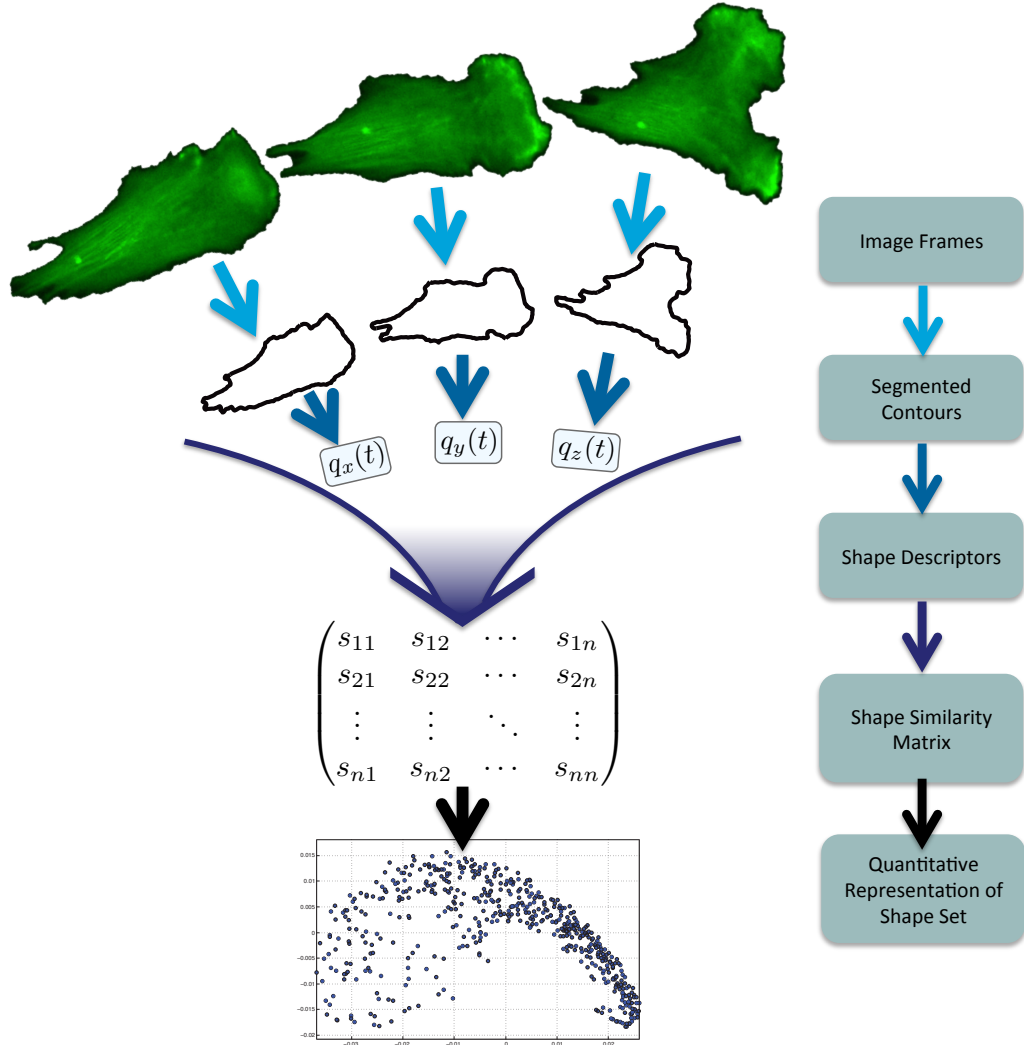


Figure 1.2: **Algorithm flow diagram.** A flow diagram illustrating the algorithm developed for generation of a low dimensional representation of cell shape. We make repeated use of this algorithm for quantitative cell shape analysis on images of migrating cells.

variance.

## 1.5 Thesis Organisation

Chapter 2 outlines all experimental techniques used to collect the data for this thesis. It will also give detail of the analytic techniques available in literature that I employed in various algorithms.

Chapter 3 gives a description of the decisions and development made to create a robust and efficient cell shape representation framework. We also present an alternative framework to independently investigate the performance of the Best Alignment Metric, a novel shape dissimilarity metric.

In chapter 4 we make use of the BAM dissimilarity metric within our cell shape representation framework for the purpose of morphological phenotyping on a large dataset of RPE cells. As shape is a quality commonly assessed subjectively, this chapter contains a number of figures that visualise the output of the shape representation framework. We also investigate the correlation between the distribution created by our framework and a number of common shape features, many of which are often linked to cell dynamics, in order to determine which of them are best represented by our framework, and therefore which are most dominant in the dynamics of RPE cells.

Chapter 5 presents an investigation into the motile behaviour of RPE cells, making use of our cell shape representation. We demonstrate that it is possible to reasonably accurately determine the location of a turn in a cell's path (through one turn mechanism at least) by examining the morphological information of the cell alone. This chapter demonstrates a successful application of the cell shape representation framework and provides evidence for a measurable relationship between cell morphology and migratory behaviour.

Chapter 6 concludes the thesis with a discussion of successes and limitations of the work within the context of current literature, and suggests possible directions for future work.

## Chapter 2

# Materials and Methods

### 2.1 Cell Culture and Imaging

#### 2.1.1 Cell Culture

Human retinal pigment epithelial (RPE1) cells immortalised with hTERT (Clontech) were grown in RPE medium (DMEM/F-12 medium containing 10% FCS, 2.3 g/l sodium bicarbonate, 2mM L-Glutamine, 100 U/ml penicillin and 100 µg/ml streptomycin) at 37°C, 5% CO<sub>2</sub> in a humidified incubator. The RPE1 GLA6 cell line was generated by transfecting hTERT RPE1 cells (Clontech) with mGFP-LifeAct [Riedl et al., 2008] followed by selection with 500 µg/ml Geneticin (Invitrogen). For depletion experiments, small interfering RNA oligonucleotides targeted against Kif1C (5-CCCAUGCCGUCUUUACCAU-[dC]-[dG]-3) or a scrambled control (5-GGACCUGGAGGUCUGCUGU-[dT]-[dT]-3) were transfected using Oligofectamine (Invitrogen) following manufacturer’s instructions. Cells were analysed 48 hours after transfection. Depletion efficiency and specificity was validated using immunofluorescence and Western blotting [Theisen et al., 2012].

#### 2.1.2 Imaging

For live cell imaging, 35mm glass-bottom dishes (Fluorodish) or 2-well chambered coverglass chambers were coated with 10 µg/ml fibronectin (Sigma). The fibronectin solution was allowed to incubate for 2-12 hours, and was washed twice with ddH<sub>2</sub>O before equilibrating the chamber with RPE medium. 6000 RPE1 GLA6 cells were seeded into each dish/well and allowed to spread for 4-6 hours. Cell migration experiments were carried out in RPE growth medium in a microscope stage top incubator (Tokai Hit) heated to 37°C and providing 5%CO<sub>2</sub>. In each experiment, numerous fields of migrating cells were imaged every 5 min for 12 hr using a 10x

objective on an Olympus personal Deltavision microscope (Applied Precision, LLC) using a GFP filter set (Chroma) and a Coolsnap HQ camera, controlled by Softworx (Applied Precision, LLC). Frame rate was set at imaging every 5 minutes because this was adequate for tracking purposes, since the cells neither moved nor changed shape suddenly over this time period, and with any faster imaging we would begin to see the cells negatively affected due to photodamage. The resulting images acquired at every time point were 1024x1024 pixels with 645nm/pixel resolution.

## 2.2 Cell Segmentation

To capture cell shape, we extract the outline of cells from image sequences of mGFP-LifeAct-labelled cells. Only those cell outlines were included in the analysis that did not touch the borders or any other cell in the images. The minimal number of consecutive frames needed for inclusion in the dataset was 5 frames. To find the cell boundary, we used a mean shift algorithm embedded into a graphical user interface. We used the Edison Matlab interface for mean shift using the following parameters: SpatialBandWidth = 5, RangeBandWidth = 3, Colour Space = LUV. Segmentation errors that resulted in the fragmentation of long cell extensions were manually fused to prevent bias in the dataset for compact cell shapes that segment more easily.

## 2.3 Diffusion Maps

The Diffusion Maps (DM) framework is a non-linear dimensionality reduction technique that generates a low-dimensional coordinate representation of data. Similar data points in the high-dimensional space are represented by new low-dimensional points that are close; dissimilar data points are represented by new low-dimensional points that are far apart.

To perform a DM based low-dimensional embedding of  $n$  contours,  $\{f_i\}$  where  $1 \leq i \leq n$ , one constructs an  $n \times n$  matrix  $P$  with its  $(j, k)$ th entry given as follows,

$$p_{jk} = \frac{w(f_j, f_k)}{\sum_{i=1}^n w(f_j, f_i)}, \quad (2.1)$$

where  $w(\cdot, \cdot)$  is the chosen shape similarity measure. This matrix  $P$  can be thought of as a Markov transition matrix (where similarity is analogous to diffusion distance). Then we perform eigen-decomposition upon  $P$ , and we know from the Perron-Frobenius theorem that  $P$  has exactly one eigenvalue equal to 1 and all

other eigenvalues have a strictly smaller magnitude. So (by reordering if necessary) let  $1 = \lambda_0 > |\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_{n-1}|$  be the set of eigenvalues, and  $\{\psi_i | i = 0, \dots, n-1\}$  be the set of corresponding  $n$ -dimensional eigenvectors. Then, if  $\psi_i^{(j)}$  is the  $j$ th component of the eigenvector  $\psi_i$ , we construct a lower dimensional representation of contour  $f_j$  as

$$\varphi_j = (\lambda_1^t \psi_1^{(j)}, \lambda_2^t \psi_2^{(j)}, \dots, \lambda_\rho^t \psi_\rho^{(j)}), \quad (2.2)$$

where  $\rho \ll n$  is our choice of dimension for the embedding, and  $t$  denotes time in the Markovian sense (we chose  $t = 1$  in our analysis, as we are interested in local geometric properties of shape space). Note that  $\rho$  is chosen to be much lower than the dimensionality of the original data, and hence  $\varphi_j$  is a low dimensional embedding of the contours. In a similar fashion to other dimensionality reduction techniques,  $|\lambda_i|$  reflects the proportion of the overall variance of the dataset that is accounted for in eigenvector  $\psi_i$ . Hence  $\rho$  can be chosen to be large enough to give the desired accuracy.

### 2.3.1 Laplace-Beltrami Normalisation

In order to deal with data that is sampled with non-uniform density it is possible to incorporate Laplace-Beltrami Normalisation within the Diffusion Maps framework [Lafon and Lee, 2006]. This is done by replacing both instances of the similarity measure  $w(.,.)$  of equation 2.1 with the normalised measure  $\tilde{w}(.,.)$  defined to be

$$\tilde{w}(f_j, f_k) = \frac{w(f_j, f_k)}{\sum_{i=1}^n w(f_j, f_i) \sum_{i=1}^n w(f_k, f_i)}. \quad (2.3)$$

## 2.4 Square-Root Elastic (SRE) distance

Joshi et al [Joshi et al., 2007] have presented a framework for consideration of shapes that is suitable for our analysis. While their framework is well defined for all absolutely continuous curves in  $\mathbb{R}^n$ , we will restrict our use to unit path-length closed curves in the plane. Given a closed curve in the plane,  $\alpha : \mathbf{S}^1 \rightarrow \mathbb{R}^2$  we look at its Square-Root Velocity representation,  $q : \mathbf{S}^1 \rightarrow \mathbb{R}^2$ , defined as

$$q(t) = \frac{\dot{\alpha}(t)}{\sqrt{\|\dot{\alpha}(t)\|}}. \quad (2.4)$$

The space of all such curves is defined as preshape space ( $\mathcal{C}$ ). Hence, the construction of  $\mathcal{C}$  is as follows:

$$\mathcal{C} = \left\{ q \in \mathbb{L}^2(\mathbf{S}^1, \mathbb{R}^n) \mid \int_{\mathbf{S}^1} \|q(t)\|^2 dt = 1, \int_{\mathbf{S}^1} q(t) \|q(t)\| dt = 0, \right\}, \quad (2.5)$$

where  $\int_{\mathbf{S}^1} \|q(t)\|^2 dt = 1$  provides the restriction to unit length and  $\int_{\mathbf{S}^1} q(t) \|q(t)\| dt = 0$  provides the restriction that the curves are closed. Then, to tackle the issue of appropriate invariances (see section 3.2.1), the authors introduce shape space ( $\mathcal{S}$ ) as the quotient of preshape space by the groups of reparameterisations ( $\Gamma$ ) and rotations in the plane ( $SO(2)$ ) i.e.  $\mathcal{S} = \mathcal{C}/(\Gamma \times SO(2))$ .

They present an algorithm [Srivastava et al., 2011] for determining geodesics in preshape space ( $\mathcal{C}$ ) that minimise path length according to the Elastic metric [Mio et al., 2007]. So the distance between any two curves  $q_0$  and  $q_1$  can be defined as

$$d_c(q_0, q_1) = \inf_{\{\kappa: [0,1] \rightarrow \mathcal{C} \mid \kappa(0)=q_0, \kappa(1)=q_1\}} L(\kappa), \quad (2.6)$$

where  $L(\kappa) = \int_0^1 \sqrt{\langle \dot{\kappa}(t), \dot{\kappa}(t) \rangle} dt$  is the length of  $\kappa$  (a path on  $\mathcal{C}$ ), according to the Elastic metric,  $\langle \cdot, \cdot \rangle$ , as defined in [Mio et al., 2007].

This is used in a second algorithm which finds the geodesic distance in shape space ( $\mathcal{S}$ ). The geodesic distance in shape space between shapes  $[q_0]$  and  $[q_1]$  is defined as

$$d_{\mathcal{S}}([q_0], [q_1]) = \inf_{\{(\gamma, \mathcal{O}) \in \Gamma \times SO(2)\}} d_c(q_0, \mathcal{O}(q_1 \circ \gamma) \sqrt{\hat{\gamma}}). \quad (2.7)$$

This we refer to as the Square-Root Elastic (SRE) distance.

## 2.5 Affinity Propagation

Affinity Propagation [Frey and Dueck, 2007] is a clustering algorithm that selects a subset of the data to be “exemplars”; all elements are then assigned to exactly one exemplar. Hence, each exemplar forms a cluster with the points that are assigned to it. The algorithm seeks to find the cluster/exemplar configuration that maximises the total sum of exemplar preferences and the similarities between points and their exemplars. This is achieved rapidly by a message passing process that iteratively passes information between nodes and updates the system.

To perform AP clustering on a dataset of size  $K$ , the algorithm requires a similarity matrix  $\{s_{ij}\}$  and a set of preferences  $c_k$ , for all  $i, j, k = 1, \dots, K$ . We set  $c_k$  to be constant over  $k$ , and equal to the median of  $\{s_{ij}\}$ . The choice of similarity measure

is application specific, we discuss our choice for shape clustering in section 3.4.3. The following messages are computed iteratively:

$$\alpha_{ij} = \begin{cases} c_j + \sum_{k \neq j} \max(0, \rho_{kj}) & i = j, \\ \min[0, c_j + \rho_{jj} + \sum_{k \notin i, j} \max(0, \rho_{kj})] & i \neq j, \end{cases} \quad (2.8)$$

$$\rho_{ij} = s_{ij} - \max_{k \neq j} (\alpha_{ik} + s_{ik}), \quad (2.9)$$

where  $\alpha_{ij} = 0$  initially. The value of  $\rho_{ij}$  can be thought of as a measure of how well suited  $j$  is as an exemplar for  $i$ , taking into account other potential exemplars for  $i$ . The value of  $\alpha_{ij}$  can be thought of as a measure of how available  $j$  is to serve as the exemplar for  $i$ , taking into consideration other points for which  $j$  is an exemplar.

## 2.6 Seriation Algorithm

This is an algorithm designed for a package for cluster analysis [Wishart, 1999], and it deals with the reordering of branches of a dendrogram in order to optimise the rank order of the corresponding similarity matrix. A dendrogram is a way of illustrating the results of hierarchical clustering [Ward, 1963], but the displayed order of the branches is not considered. In fact there are  $2^{n-2}$  ways of rearranging a dendrogram of  $n$  elements.

The seriation algorithm chooses an order that optimises the rank order of the similarity matrix, which is a matrix with elements  $s_{ij}$  equal to the similarity between elements  $i$  and  $j$ . We construct a matrix  $A$ , corresponding to the rank of each element in a row, i.e. in row  $i$  let  $a_{ii} = 0$  and  $a_{ik} = 1$  where  $k \neq i$  is the index of the element most similar to element  $i$ , and  $a_{im} = 2$  where  $m \neq i$  is the index of the element next most similar to  $i$  and so on. The goal is then to rearrange the rows and columns (symmetrically) to make this rank matrix as close as possible to the perfect rank matrix:

$$\begin{array}{cccccc} 0 & 1 & 2 & 3 & \dots & \\ 1 & 0 & 1 & 2 & \dots & \\ 2 & 1 & 0 & 1 & \dots & \\ 3 & 2 & 1 & 0 & \dots & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \end{array} \quad (2.10)$$

Specifically the algorithm tries to minimise the value of

$$\rho = 1 - \frac{\sum_i \sum_j (a_{ij} - p_{ij})^2}{(n^3 - n)}, \quad (2.11)$$

where  $a_{ij}$  are the row-wise rank elements as before and  $p_{ij}$  are the corresponding perfect rank matrix elements. Full details of the optimisation procedure can be found in [Wishart, 1999].

## 2.7 Hidden Markov Models

In Chapter 5 we use Hidden Markov Models to predict cell behaviour from cell shape information [Baum and Petrie, 1966]. We considered four hidden states of cell morphology: a depolarised state, a polarised state and the two transition states: depolarising and repolarising.

Hidden Markov Models are used to represent a situation involving hidden states that govern some observable variables. Normally one hidden state will be considered active at a given time. The state will have an emission distribution governing the observables, and transition probabilities that determine the probability that each of the hidden states will be active at the next time step.

If the emission distributions and transition probabilities are known for each state, and we are given a sequence of observed variables, often the challenge is to find the most likely sequence of states to have produced these emissions. To implement the Hidden Markov Model, we used the Probabilistic Modelling Toolkit version 3 [Murphy and Dunham]. This toolbox uses of the Viterbi algorithm [Viterbi, 1967] (summarised below) to find the most likely states for our data sequences.

Assume we have an observed data sequence  $x_1, x_2, \dots, x_T$ , transition probabilities  $\tau_{i,j}$  from state  $i$  to state  $j$  in state space  $S$ , initial probabilities  $\pi_i$  of being in state  $i$  at time 0. Then define the following values:

$$V_{1,k} = P(x_1|k) \cdot \pi_k, \quad (2.12)$$

$$V_{t,k} = P(x_t|k) \cdot \operatorname{argmax}_{s \in S} (\tau_{s,k} \cdot V_{t-1,k}). \quad (2.13)$$

Here  $V_{t,k}$  represent the probability of being in state  $k$  given the observed variables up until time  $t$ . Then the most likely states for each time are determined in reverse



order as follows

$$q_T = \operatorname{argmax}_{k \in S} V_{T,k} \quad (2.14)$$

$$q_t = \operatorname{argmax}_{k \in S} (\tau_{k,q_{t+1}} \cdot V_{t,k}), \text{ for } t = 1, \dots, T - 1. \quad (2.15)$$

We trained the model using 19 manually selected image sequences that were deemed typical examples of a turn through depolarisation/repolarisation. The four states were classified manually and the distribution of the shapes in our shape matrix determined.

## 2.8 Cell Track Data

In chapter 5 we investigate the migrational behaviour of RPE cells. In order to separate analysis of the migration from analysis of the shape of the cells we define a cell's track to be the path of the centroid of the cell over the course of the image sequences. The centroid is simply computed as the mean of all pixel positions contained within the cell segmentation. This is computed through the use of the MATLAB function 'regionprops'.

## Chapter 3

# The Best Alignment Metric

### 3.1 Introduction

Our proposed framework for shape analysis makes use of the Diffusion Maps algorithm for manifold learning. Applying this algorithm to a given dataset requires the selection of a suitable similarity measure. We make use of the simple Gaussian kernel for data points  $x$  and  $y$ ,

$$w(x, y) = \exp\left(\frac{-d(x, y)^2}{2\sigma^2}\right), \quad (3.1)$$

where  $d(x, y)$  corresponds to a chosen distance metric (for us this will be a difference measure between shapes  $x$  and  $y$ ) and  $\sigma$  corresponds to a chosen kernel bandwidth. The main focus of this chapter is the development of the Best Alignment Metric (BAM), which is our chosen shape distance metric. We will also touch on the choice of kernel bandwidth.

### 3.2 Shape Difference Metric

#### 3.2.1 Understanding the data

Section 2.2 describes our protocol for extracting a cell's outline from a cell image. We choose to use cell outlines to represent cell shape to remove other information about the cell's intracellular activity and just focus on shape. But there are still pieces of extraneous information in this representation that our analysis needs to be invariant to; namely the cell's position, its angular orientation, and the parameterisation of the cell outline. However the term *invariant* is somewhat of a red herring, since some methods of introducing invariance also introduce significant errors in our

framework. To understand this issue, it is important to note that we are comparing a *pair* of shapes, and that some information is extraneous to the individual cells, but significant in a relative sense. Specifically, the position and orientation of a cell in a frame does not matter, but to compare two cells we must be careful to control their mutual alignment. One big pitfall here is that many of the common solutions to invariance do not do this. One common style of rotationally invariant shape representation, we call it the standard form, will consistently represent shapes as if they were in a particular orientation, and so will be rotationally invariant but will give no consideration to whether any pair of represented shapes is appropriately mutually aligned. This problem is obvious when presented in figures such as figure 3.1, but the same problem is less clear (but still present) in methods using chain code or Fourier representation, for example.

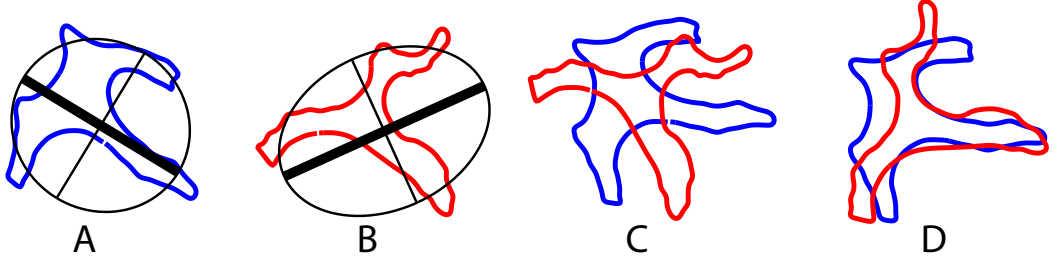


Figure 3.1: **Cell shape misalignment.** This figure illustrates the difficulties faced when mutually aligning complex shapes along intrinsic axes. The curves labelled A & B show cell contours and their best-fit ellipses with thick major axes. C shows the result of aligning the two major axes (a common approach). D shows a more suitable alignment of the two shapes. Figure reproduced from [Jefferyes et al., 2013].

The simplest solution is to remove this information entirely, for example if we simply compare the length of each cell we would not need to worry about their mutual alignment, however we obviously want to include a lot more information than this. The difficulties in selecting a representation that truly does not carry our identified extraneous information but does, however, carry all other information and additionally creating a similarity measure that appropriately reflects shape similarity seemed vast, so we opted for another strategy. The strategy that we decided would be most reliable was to mutually align each pair of shapes.

In section 2.4 we outline the Square Root Velocity representation [Srivastava et al., 2011] and their shape distance framework involving an Elastic Metric (first introduced in [Younes, 1998]). This framework is free of the requirements for shape landmarks, provides an invariant representation, and is also highly regarded as an intuitive representation of the structure of shape space. As such it fulfils our cri-

teria for appropriate handling of the data. We made extensive use of this metric in our preliminary work and employed its use in our publication [Jefferyes et al., 2013]. We found in our early experiments that the datasets were too limited and did not reflect the full dynamic range of cellular behaviour. We needed to increase the size of our dataset and required an algorithm that could handle larger datasets, however the complexity of the SRV algorithm means it was prohibitively slow for these requirements.

This led us to the formulation of the Best Alignment Metric (discussed in section 3.2.2). Beneath this metric is a very simple notion of curve distance. However rather than only considering curves, we consider equivalence classes of curves, i.e. the set of all curves that only vary through operations we would consider irrelevant. We make use of circular convolution in Fourier space to rapidly find the optimum choice amongst all possible pairwise matchings between equivalence classes. We performed some experiments to show that results are comparable to the SRV framework, but computation time is dramatically lower, and so we brought the BAM algorithm forward to incorporate into the Diffusion Maps framework for shape representation.

### 3.2.2 Introducing the Best Alignment Metric

In this section we give an overview of the theory and motivation behind the Best Alignment Metric; for a full brute force proof, see appendix A. At its heart, BAM is based simply on the  $L^2$ -norm between curves;

$$\|u - v\|_{L^2}^2 = \int \|u(s) - v(s)\|^2 ds \quad (3.2)$$

where  $u, v \in \mathcal{C}^\infty([0, 1], \mathbb{R}^2)$  are curves in the plane. In our case these are simple closed curves ( $u(0) = u(1)$ ,  $u(a) \neq u(b)$  for  $a, b \in (0, 1)$ ,  $a \neq b$ , equivalently for  $v$ ).

Here, the word *curve* is used in reference to an explicit curve in the plane. It is important to distinguish a *curve* drawn in the plane from a *shape* that is independent from a coordinate system. This notion of *shape* is formally defined as an equivalence class of curves over the standard operations of translation, rotation and cyclic reparameterisation.

The BAM distance is defined over these equivalence classes. For a given curve,  $u$ , we denote the equivalence class as  $[u]$  and define the BAM distance as

$$d_{BAM}([u], [v])^2 = \min_{(r, \theta)} \int \|v_t(s) - \text{rot}_\theta(u_t(s + r))\|^2 ds \quad (3.3)$$

where the argument  $(s + r)$  is taken modulo 1, the minimum is taken over  $[0, 1) \times [0, 2\pi)$ ,  $rot_\theta$  represents a planar rotation of angle  $\theta$  centred at the origin, and  $u_t$  (resp.  $v_t$ ) represents the curve  $u$  (resp.  $v$ ) translated so that the mean of the curve lies on the origin. In future, we will omit this subscript and all curves can be assumed to lie with their mean at the origin. In the appendix A.2, we provide proof that this translation to the origin minimises the  $L^2$ -norm over all other possible translations.

The above definition is presented for continuous curves. However, in practise of course, the boundaries of cells are observed and represented as discrete approximations. We therefore redefine BAM for discrete curves  $u = \{u_j \in \mathbb{C} : j = 0, \dots, N-1\}$  (note that the number,  $N$ , of points used to represent each curve must be fixed across the dataset and points must be evenly spaced around the curve). BAM for discretely represented curves is defined as

$$d_{BAM}([u], [v])^2 = \frac{1}{N} \min_{(r, \theta)} \sum_{j=0}^{N-1} |v_j - e^{i\theta}(u_{j+r})|^2. \quad (3.4)$$

Here (as later) the index  $(j + r)$  is taken modulo  $N$ , and the minimum is taken over  $\{0, \dots, N-1\} \times [0, 2\pi)$ . As discussed earlier, this is a very simple and intuitive measure of shape difference. Its power comes from its ability to be reformulated to the following expression, which admits a very rapid implementation

$$Nd_{BAM}([u], [v])^2 = \sum_{j=0}^{N-1} |v_j|^2 + \sum_{j=0}^{N-1} |u_j|^2 - 2 \max_r \sum_{j=0}^{N-1} |v_j \overline{u_{j+r}}|. \quad (3.5)$$

The reasons that this admits a rapid implementation are threefold. Firstly, many of the terms depend only on one curve and so can be computed only once per curve, not per pair of curves. Secondly,  $\theta$  is removed, as this formulation explicitly computes the appropriate quantity over all rotations. Thirdly, the last term in the expression can be rapidly computed through use of circular convolution. For a brute-force style formulation and proof of BAM and the claims above, see the appendix A.

Algorithm 1 presents the pseudocode for measuring BAM over a large dataset of curves. It highlights the fact that many of the terms can be calculated once for each curve, rather than each pair of curves, which saves a large calculation cost.

---

**Algorithm 1** Computing the Best Alignment Metric between pairs of curves in a large dataset.

---

**Input:**  $\mathcal{U}$ , a set of  $M$  planar curves. Each curve is represented by a cyclic sequence,  $u = (u_0, u_2, \dots, u_{N-1})$ , of  $N$  equally spaced complex numbers with mean equal to zero.

**Output:**  $D$ , an  $N \times N$  dissimilarity matrix.

*for-loop* over  $u \in \mathcal{U}$

1. Compute  $s(u) = \sum_{i=0}^{N-1} |u_i|^2$ .
2. Compute  $(c_{u,j})_{j=0}^{N-1}$ , the fast Fourier transform of  $(\overline{u_j})_{j=0}^{N-1}$ .
3. Compute  $(f_{u,j})_{j=0}^{N-1}$ , the fast Fourier transform of  $(u_{(N-j-1)})_{j=0}^{N-1}$ .

*end*

*for-loop* over  $u \in \mathcal{U}$

*for-loop* over  $v \in \mathcal{U}$

1. Compute  $(X_j)_{j=0}^{N-1}$ , the inverse fast Fourier transform of  $(c_{u,j} f_{v,j})_{j=0}^{N-1}$ .
2. Compute  $A = \max_j |X_j|$ .
3. Compute  $D(u, v) = \sqrt{s(u) + s(v) - 2A}$ .

*end*

*end*

---

### 3.2.3 Discussion of the Best Alignment Metric

Designing an efficient and effective difference metric is a common challenge in computer science. The Best Alignment Metric (BAM) is extremely fast and we believe it will be useful to others working in shape comparison. However BAM has been designed with our application in mind and it may not be suitable for all applications. In this section we discuss certain features of BAM that potential users need to consider.

The first issue to consider is that when comparing two shapes that are the mirror images of each other, BAM will produce a non-zero score (unless obviously, the shapes are identical because they are symmetric). This is suitable for our work because we consider it an interesting difference in cells, for example, it maybe useful in distinguishing a cell turning left from a cell turning right. For this reason, in an application where shape similarity ought to be invariant to reflection, BAM would not be appropriate without alteration.

Another issue involves whether or not to standardise the scale of the shapes being analysed. In many applications of shape analysis, scale invariance is included because the shape of an object can be considered independent to its scale. In our case

there is an argument that cells of different sizes may undergo similar morphological changes, e.g. bending or protruding, and so standardising the scale of the shapes allows us to more accurately measure their similarity. Another argument is that in our context, since the microscope is always at a fixed distance from the cells and we standardise the magnification, any difference in the size of the cells is a genuine biological difference between the cells and this may be significant. We opted to make our analysis scale invariant by standardising the perimeter of our shapes. However this turned out to be relatively inconsequential; when we investigated how simple shape features distributed across our low-dimensional representation (as discussed later in section 4.2.3) scale related features such as area and perimeter seemed remarkably ordered in the low-d space as compared to orientation (compare figures 4.3 and 4.11 to figure 4.7). This suggests that size is relatively conserved in our cells and changes in size only occur with other morphological changes.

### 3.2.4 BAM comparisons

In this section we discuss the performance of BAM, comparing it to other approaches. We measure the speed of some shape distance measures. To do this, we measured the total time taken to compute the distances of 100 pairs, dividing that number by 100 to give a measure of the time to compute the BAM distance once. We ran this process 100 times to create a mean and standard deviation. We calculated that the average time to compute one BAM measurement is  $35 \pm 1.6 \mu\text{s}$  (microseconds). Below we compare BAM to two other methods (Symmetric Difference and Fourier Descriptors) and discuss the use of Shape Features.

#### Symmetric Difference

A mathematical interpretation of the value of BAM is the integral of the symmetric difference of the two shapes minimised over rotation and translation. We can attempt to compute this value more literally by calculating the size of the symmetric difference (using the XOR logical operation over matrices) between binary images. We calculate this difference between one target image and a range of rotated images and take the minimum value. The speed and accuracy of this calculation will obviously depend on your choices for the resolution of the binary images and the number of angle options calculated. At a resolution of  $128 \times 128$  pixels and 10 rotation options, we calculated that the average time to compute one symmetric difference measurement is  $53 \pm 9.2 \text{ ms}$  (milliseconds). Even at this low resolution, computation time is already 1500 times slower than BAM. The result of the com-

putation cost difference is that for a dataset with 10,000 samples computing the pairwise distances with Symmetric Difference would take over a month, whereas with BAM this process would take under an hour. In our analysis of RPE cells we examine a dataset of nearly 38000 cell shapes so this cost is significant.

Comparative performance can be seen in figure 3.2. Subjectively, it seems that there is similar performance between BAM and Symmetric Difference, there is certainly no grounds to justify the extra computation cost.

### Fourier Descriptors

A feature set that is commonly used for shape description is Fourier Descriptors. Fourier Descriptors can be generated from the cell outline by performing a Fourier transform, rotation invariance can then be gained by taking the absolute value of the Fourier transform (commonly known as the power spectrum). With the curve represented by a discrete sequence  $\{x_n\}$ , we can make use of the fast Fourier transform

$$X_k = \sum_{n=0}^{N-1} x_n e^{-2\pi i k n / N}, \quad (3.6)$$

from which we can rapidly compute the power spectrum:

$$P_k = X_k \cdot X_k^*. \quad (3.7)$$

These features represent the levels of auto-correlation at different frequencies around the cell’s edge. Simply taking the Euclidean distance in this feature space gives us a shape similarity measure. Time experiments reported an average time to compute one FD distance measurement as  $7.8 \pm 5.4 \mu s$ , making it approximately 4.5 times faster than BAM. Figure 3.2 shows the performance of this as a shape similarity measure. For the most part, Fourier Descriptors give perceptually very similar to the performance of BAM. However BAM can be seen (albeit subjectively) to outperform Fourier in rows 5, 8 and 10, at least. This emphasises the fact that aspects of shape information are lost when the phase information is removed, and the power spectrum alone is not enough to faithfully represent shape.

BAM has another advantage, in that it is defined directly on shape space. One goal for our framework is that it could be extended to allow for generation of synthetic contours from arbitrary points in our low-dimensional representation. BAM allows this, since any reversal of an embedding based on BAM maps back into shape space, whereas reversing an embedding based on Fourier Descriptors maps into the space



































































































































































		Best Alignment Metric					Symmetric Difference					Fourier Descriptors				
		1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th
1																
2																
3																
4																
5																
6																
7																
8																
9																
10																

Figure 3.2: **Comparative performance of shape distance measures.** Each row above shows a randomly selected target RPE cell outline, followed by 5 outlines identified as closest to the target outline (excluding outlines of the same cell as the target) according to 3 different shape distance measures. The difference metrics are the Best Alignment Metric (as defined in section 3.2.2), the Symmetric Difference and Fourier Descriptors (as described in Section 3.2.4).

of power spectra and since the power spectra are missing the phase information, any map back into shape space is not unique.

## Earth Mover’s Distance

The Earth Mover’s Distance [Rubner et al., 1998] is an approach used by many in shape analysis. This distance measure computes the difference between two images by calculating the amount of work required to change one image into the other. The analogy goes that one image can be seen as piles of dirt (where the height of a pile corresponds to the pixel intensity), the other as holes in the ground (depth corresponding to pixel intensity). Then if the dirt were laid on top of the holes and the images were identical the holes would fill up perfectly, otherwise, the work required to move the dirt into the holes measures the difference. This method is seen as intuitive and is popular in shape analysis. However, it is not immediately rotation invariant, this invariance must be introduced.

One method for introducing rotation invariance is to convert the images first into a rotation invariant representation, e.g. Fourier Descriptors or Shape Context [Belongie et al., 2002; Grauman and Darrell, 2004]. Here we run into the same difficulties that we discussed for Fourier Descriptors above, in that to generate these representations we must lose some information.

Another way to introduce invariance would be to pre-align the cell images. We propose that the best method for pre-aligning the images would be to actually use BAM (a discussion of alignment of contours is given in section 3.2.1).

## Shape Features

A common approach to biological analysis is to focus investigation on features that are known to be particularly important in a given situation. In shape analysis it is no different and with sufficient knowledge and understanding of the system it would be possible to design an incredibly efficient and effective way of differentiating shapes for any given task. However, as we discussed in section 1.4 we wish to develop a framework that can be applied to a situation without any *a priori* information.

## 3.3 Kernel Bandwidth

The selection of kernel bandwidth ( $\sigma$  in equation 3.1) is a very important choice in this context and appears in many other contexts in computer science, in fact optimal selection of this parameter is an active area of research. The interpretation of this parameter is contextual scale, i.e. at what distance would one say something was close and what would one call far away. Humans are very good at naturally estimating this parameter and can easily describe things as big or small, many or

few, high or low etc. even with a minimal familiarity with the distribution, but in machine vision, becoming context aware takes a lot more effort. Throughout the project we looked at a number of methods for tuning this parameter.

One method we looked at is called Reverse Soft K-Nearest Neighbour Density Estimation (RSKNNDE) [Kursun, 2010]. This was designed with spectral clustering in mind, as spectral clustering is sensitive to outliers. The method seeks to mitigate the impact of outliers by weighting each point’s contribution to the kernel estimation relative to it’s own density amongst its K-nearest neighbours. The weighted average of the density estimations is then used to create a kernel estimation.

Another method we looked at using involved a self-tuning kernel [Zelnik-Manor and Perona, 2005]. This method chooses a unique half-kernel for each datapoint, that is simply its distance from its  $K$ th nearest neighbour, then  $\sigma^2$  is set to the product of these two. The advantage here is that each distance is considered in the context that it’s in.

The method we eventually chose to use was very simple; just the median of all pairwise distances, as recommended in [Schclar, 2008]. This has been shown to be robust to outliers.

To compare the options for kernel bandwidth we can examine the performance of the downstream analysis. As a reminder, we are using this similarity kernel as part of a framework for low dimensional shape representation. The main requirements for this shape space representation are 1) similar shapes are represented by points that are close together and different shapes are represented by points far apart and 2) the important morphological features (i.e. those features that are responsible for the most variation within the dataset) are well-organised in relation to the internal structure of the point cloud in the representation space. The first requirement is mostly dependent on the choice of distance measure, and BAM performs very well in this regard. The choice of kernel bandwidth is most likely to impact on the second requirement. The ideal scenario for the second requirement is when the internal structure of the point cloud is linear and the internal axes align with the Cartesian axes of the representation space. One way in which the representation can under-perform is by the internal axes of the point cloud being curved rather than linear, in the extreme, the point cloud can be so curved that the extremities are too close in Euclidean distance. The bottom plot of figure 1.2 shows a early embedding result using the RSKNNDE method for selecting the kernel bandwidth where this phenomenon is clearly visible. For some time we thought that this curved structure might reflect some truth about the structure of shape space, however we now believe that it is an effect of the kernel bandwidth selection.

We found that selecting the kernel bandwidth using the median or using the self-tuning kernel method will both produce a point cloud with linear structure in the first two new dimensions. In the third dimension, the point cloud using the median is still linear whereas the point cloud generated using the self-tuning method is curved, so, although we do not make use of the third dimension we decided to move forward using the median method.

Our final distribution proved to be quite robust to the choice of kernel, probably because of the size of the dataset. When analysing smaller datasets we recommend careful consideration of this parameter.

## 3.4 Examining the Performance of the Best Alignment Metric

### 3.4.1 Affinity Propagation for Independent Validation

Our intention for this chapter is to find a shape metric to use within the Diffusion Maps framework as outlined in section 1.4. This framework will create a shape representation that we will later use for exploratory analysis, so it is important at this stage to attempt to validate our shape metric within a more restricted framework. To this end, we make use of Affinity Propagation clustering (see section 2.5 or [Frey and Dueck, 2007]). This algorithm is well suited for testing our metric since it's required inputs are only a similarity matrix and a set of preferences (which we compute from the similarity scores), then we can assess the validity of the cluster assignments.

In section 3.4.2 we look at the comparative performance between the Best Alignment Metric (BAM) and the Square-Root Elastic (SRE) distance (see section 2.4 or [Joshi et al., 2007]). Due to the speed restrictions of the SRE computation, this analysis looks only at relatively small datasets. In section 3.4.3 we apply BAM based AP to a much larger dataset and develop an extension to AP to cope with the difficulties of the larger dataset. We will also discuss how this framework may be used as an alternative framework for shape representation in its own right, and also (as seen in section 4.1) as a useful tool for visualising the structure captured by DM.

### 3.4.2 SRE versus BAM

The motivation behind development of BAM was to create a rapidly computable metric that allows for analysis of enormous datasets. Comparison of computation time will be presented later in this section. But it is of paramount importance that

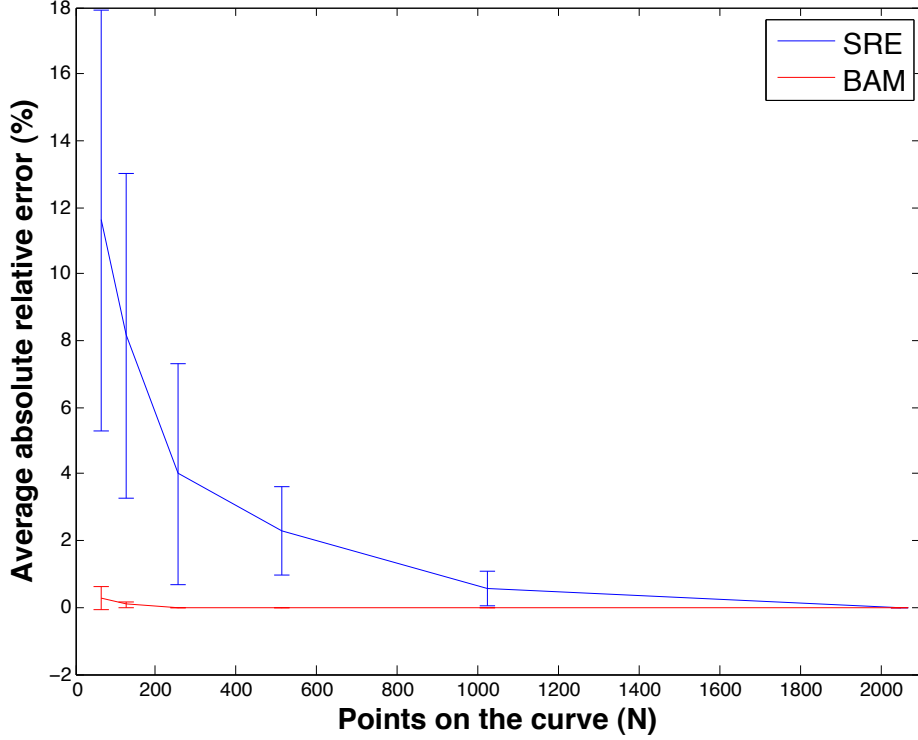


Figure 3.3: **Absolute relative error from curve sub-sampling.** A plot of the mean absolute relative error (%) of SRE (blue) and BAM (red) as functions of  $N$ , the number of points taken around the curve. Error bars show standard deviation.

the results that are computed using BAM are of high quality. The SRE distance is highly regarded as a sensible measure of shape distance. This section presents results to show that measurements produced by BAM are comparable to measurements produced in the SRE framework.

### Number of Points Around the Curve ( $N$ )

When computing distances between these shapes, something that affects both accuracy and speed is the choice of number of points around the curve ( $N$ ). An experiment was performed to examine this effect, and inform our sampling choice for later experiments. Five pairs of curves were chosen from our bank of RPE1 cell curves and the BAM and SRE distances were computed between each pair at  $N = 64, 128, 256, 512, 1024$  and  $2048$ . For each pair of shapes  $(f_i, g_i)$  for  $i = 1, \dots, 5$

$N$	Average BAM time (seconds)	Average SRE time (seconds)
64	$5.0 \times 10^{-3}$	$1.2 \times 10^0$
128	$8.6 \times 10^{-5}$	$2.9 \times 10^0$
256	$9.1 \times 10^{-5}$	$1.7 \times 10^1$
512	$1.3 \times 10^{-4}$	$1.2 \times 10^2$
1024	$1.8 \times 10^{-4}$	$9.3 \times 10^2$
2048	$3.0 \times 10^{-4}$	$7.4 \times 10^3$

Table 3.1: A table displaying the average time taken to compute SRE and BAM measurements of pairs of shapes, for a range of values for  $N$ , which is the number of points sampled from around each curve. Each result is the average measurement computed over 5 distinct pairs of shapes.

an absolute relative error was computed. We define absolute relative error as

$$ARE_i(N) = \left| \frac{(D_i(N) - D_i(2048))}{D_i(2048)} \right|, \quad (3.8)$$

where  $D_i(N)$  is either the BAM or SRE distance measured between shapes  $f_i$  and  $g_i$  at curve sample rate  $N$ . The motivation behind equation 3.8 is that accuracy will increase as  $N$  is increased, so  $D_i(2048)$  is the most accurate of our measurements. Figure 3.3 shows the mean  $ARE_i(N)$  computed over  $i$  at each  $N$ , displayed as a percentage. Here the red line represents the mean absolute relative error of BAM and the blue line represents the mean absolute relative error of SRE measurements. Error bars also show the standard deviation for each  $N$ . The error of the SRE measurements are not reliably below 1% until  $N = 1024$ , for this reason further analysis was performed on shapes represent with 1024 points.

The time taken to compute each shape distance was recorded. Table (3.1) presents the average time to compute BAM and SRE at each of the values used for  $N$ . It can be seen that at  $N = 128$ , BAM is 5 orders of magnitude faster than SRE, and at  $N = 2048$ , BAM is 7 orders of magnitude faster.

### Metric Comparison

Computing BAM is very fast, but it is very important that the resulting measure of shape distance is relevant to any given investigation. The SRE distance is a highly regarded measure of shape distance because of its intuitive theoretical construction. In this section we present a quantitative and qualitative comparison between SRE and BAM to show that the results of BAM are comparable to that of SRE. Four datasets of curves were used in this analysis. Three datasets were drawn from our

stock dataset of boundaries of migrating epithelial cells. The same experiment was then performed on data from a dataset used widely in computer vision literature. The fourth dataset was small subset of the MPEG-7 core experiment (CE) Shape-1 Part-B dataset [Latecki], which comprised of 5 randomly selected members of each of the following classes: `spoon`, `apple`, `heart`, `bat` and `chicken`.

For each of these datasets the SRE and BAM distance between every distinct pair of shapes was computed. As mentioned above we wished to assess both the quantitative and qualitative comparability of the shape distances. The quantitative analysis consisted of examining the correlation statistics between the pairwise distances. For qualitative analysis we felt it was important to assess how similarly the two metrics worked in application. Hence the distances were used to create a similarity matrix (using equation 3.1) for each dataset, which was used to perform Affinity Propagation clustering Frey and Dueck [2007]. Results can be seen in Figure 3.4, and are discussed below. Figure 3.4 contains four scatter plots showing pairwise SRE and BAM measurements for each dataset. All plots seem to indicate a positive correlation between the two distance measures.

The Spearman’s rank correlation coefficient between the SRE and BAM measurements was computed on each dataset, this had a mean and standard deviation of  $0.71 \pm 0.85$ . High values of this correlation coefficient suggest that there that the two distance measurements are likely to be related by a monotonic function. There is clearly a high standard deviation and so we cannot argue that correlation is strong in a numerical sense. But we can still examine performance to see if the output is qualitatively similar.

Figure 3.4 contains an array of shapes for each of the four datasets. These shape arrays present the results of affinity propagation clustering according to both SRE and BAM for each dataset. In each array, shapes are separated vertically by cluster assignment according to SRE. That is to say, there is no row that contains shapes from more than one SRE cluster (although some larger clusters span more than one row). The black shapes in the first column of each array are the exemplary shapes according to SRE, these are aligned with the first row of each cluster. To the right (and sometimes below) each exemplar, the whole cluster (including the exemplar again) is displayed and coloured according to BAM based cluster assignment. If the clustering assignments match perfectly, each row is coloured one unique colour. In discussion, SRE clusters shall be referred to as numbered clusters from top to bottom, BAM clusters shall be referred to by their colours.

In RPE Set 1 it seems that both clustering assignments successfully separate longer shapes that have pronounced tails from the smaller rounder shapes. However, these

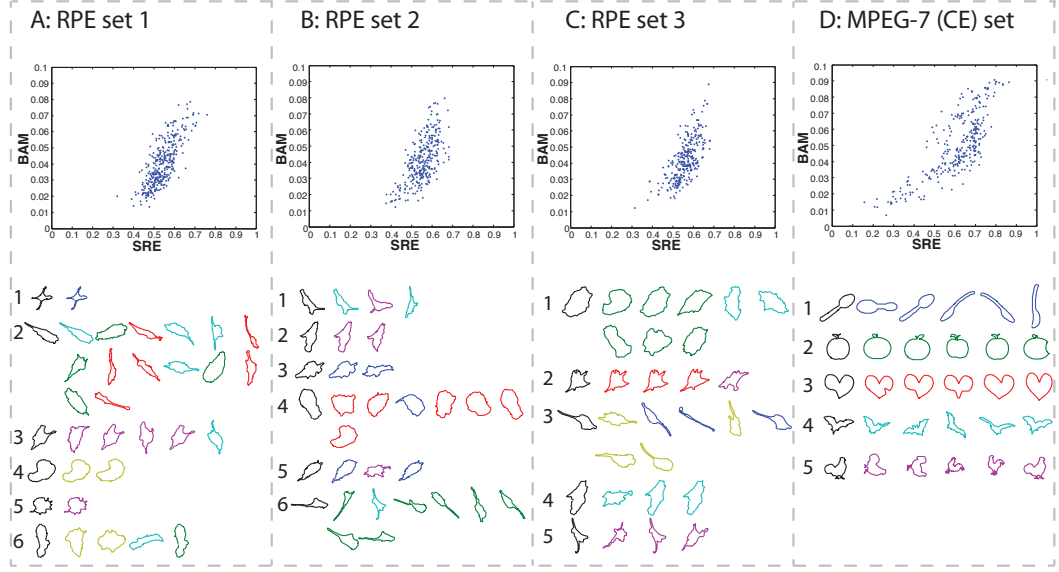


Figure 3.4: **Metric Comparison.** A figure showing quantitative and qualitative comparisons of the measurements of SRE and BAM on four different datasets. Scatter plots display the SRE and BAM pairwise distances, demonstrating the extent of the correlation between the distances. The arrays of shapes show the results of affinity propagation, which was separately performed on the SRE and BAM based similarity matrices of each dataset. The shapes are separated vertically by cluster according to SRE. To the left of the first row of each cluster the exemplary shape of that cluster is repeated in black. The (non-black) shapes are coloured according to their cluster assignment by BAM. If the clustering assignments match perfectly, each row is coloured one unique colour. Datasets A, B and C are disjoint sets drawn randomly from a larger dataset of RPE cell outlines. Dataset D is a set of 25 shapes drawn from 5 classes of the MPEG-7 (CE) dataset.

longer shapes make up one large cluster in the SRE based assignment (cluster 2), but are distributed into three smaller clusters in the BAM assignment (cyan, red and green). Both clusterings successfully identify the cross shaped cell contour as an outlier and not part of any other class. Cluster 3 matches the magenta cluster in all but one shape each. In RPE Set 2 it appears that clusters 4 and 6 match the red and green clusters (respectively) quite well. Clusters 4 and 6 contain 1 extra cell each, and the shapes of these extra cells are arguably similar to the other cells in the clusters. These clusters seem to represent the most dominant phenotypes in the dataset, long and thin versus short and round. RPE Set 3 has quite good agreement between the clustering assignments, with cluster correspondence of 1 to green, 2 to red, 3 to yellow and blue, 4 to cyan and 5 to magenta. The yellow and blue clusters arguably correspond to subclasses of the phenotype identified in cluster



3. Other small discrepancies seem to involve cells that could arguably lie in either of the identified clusters.

Affinity propagation upon the MPEG-7 (CE) set worked equally well with both shape measures, in that both measures resulted in perfect clustering.

### 3.4.3 Affinity Propagation on a Large Dataset

We now examine the performance of BAM alone, on a much larger dataset, and try to simply visually present the clustering assignment. Figure 3.5 displays the 485 exemplars generated from running AP on a dataset of 37818 RPE cell shapes (see section 2.1 for experimental information), as well as some randomly chosen elements from 6 randomly chosen clusters. A perfect clustering algorithm will have low intra-cluster variation and high inter-cluster variation. It is hopefully apparent<sup>1</sup> that the clusters have good (low) intra-cluster variation, in that the shapes in each cluster appear to share features. However it is also clear that inter-cluster variation is not consistently high, i.e. some of the exemplars are very similar to each other. This could be seen as redundancy in the model, to be tuned away with more appropriate parameters, however I believe this actually reflects the continuity of the data, i.e. hard clustering is not an accurate representation of the data. The cluster boundaries may therefore be an artifact of the algorithm, at some point the training just fits to the noise because there are no hard boundaries to be found. With that in mind, it must be remembered that the clusters are not independent from their neighbours, so we have developed an extension which attempts to incorporate this structure and gives us options for more quantitative models.

### 3.4.4 Seriation Extension to Affinity Propagation

In this section we outline an extension to Affinity Propagation that makes use of hierarchical clustering [Ward, 1963] and Wishart seriation (see 2.6 and [Wishart, 1999]). This is to overcome the issue that the hard clustering produced by Affinity Propagation is inappropriate for our continuous shape data. The intention is to measure the inter-cluster similarity and re-order the exemplars to best preserve the continuity in the data.

The result of Affinity Propagation is an assignment of each datapoint to an exemplar point. If we call this exemplar list  $\mathcal{J}$ , and we recall the shape similarity matrix,  $\{s_{ij} : 1 \leq i, j \leq K\}$ , we can examine the submatrix constructed by selecting only

---

<sup>1</sup>N.b. There is no ground truth with which to judge similarity of cell shapes. We instead ask the reader to use their visual perception to assess the fidelity of our clusters. While visual perception is obviously subjective, it is arguably closest to the truth in the case of shape similarity.

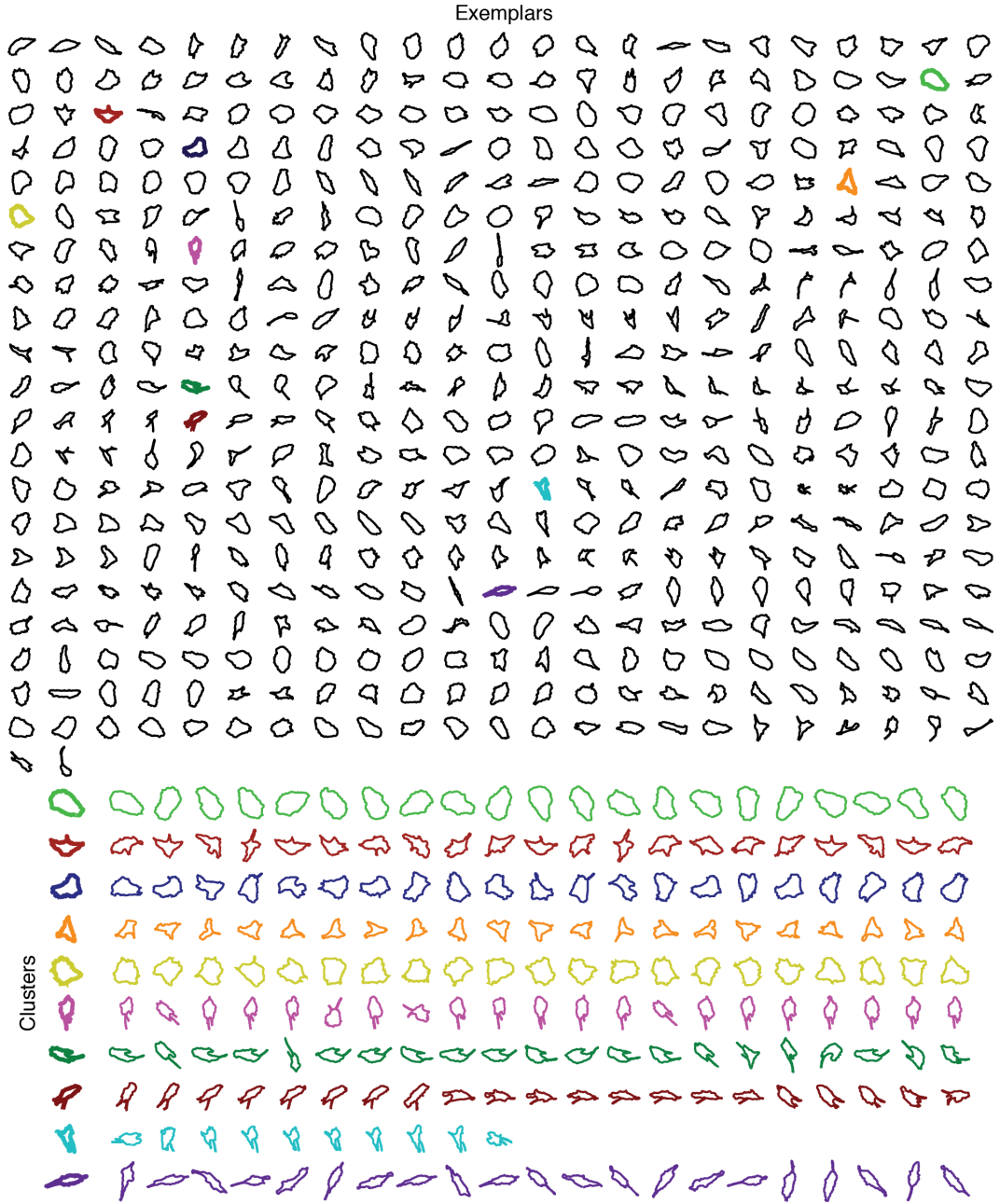


Figure 3.5: **Affinity Propagation Exemplars.** This figure shows the result of Affinity Propagation clustering on 37818 RPE1 cell shapes. At the top of the figure we show the 485 exemplars that result from AP clustering. Ten clusters were randomly chosen, and 21 elements (where available) were randomly chosen from each cluster to demonstrate the fidelity of the cluster assignment. We highlight the fact that while intra-cluster variation is low in this cluster model, inter-cluster variation is also low so the model has redundancy.

the columns corresponding to exemplars, i.e.  $\{s_{ij} : 1 \leq i \leq K, j \in \mathcal{J}\}$ . This  $K \times |\mathcal{J}|$  submatrix, which we will now call  $\hat{\mathcal{S}}$ , can be used to look at the inter-cluster similarity, since any two exemplars that are themselves similar, should produce similar similarity scores with respect to any given third shape. We therefore look at the correlation matrix,  $\mathcal{C}$ , of  $\hat{\mathcal{S}}$ , with the notion that a high correlation score will mean two clusters are similar and a low correlation score will mean they're different. This correlation matrix can be treated as a similarity matrix for the clusters and we can apply seriation as described in 2.6. This algorithm is used to reorder the rows and columns of the correlation matrix to best reflect row-rank order, we can then put the exemplars into the same order as these rows and columns to create an ordered exemplar list.

Figure 3.6 shows the exemplars again, this time following the ordered exemplar list (presented left to right, line by line). It is clear again that the 485-cluster model has a lot of redundancy, in that many of the exemplars have very similar qualities and should arguably be grouped into the same cluster. In fact it is even clearer here since similar exemplars are placed next to each other. However in this figure it is also possible to see that short sequences of exemplars show gradual progressive changes, highlighting the continuity of the data.

There are noticeable points of discontinuity, simply because the data here is forced into a 1-dimensional format in this list and this is the best solution.

The seriation process is designed to reorder the branches of a dendrogram produced by hierarchical clustering. To produce the new exemplar order, we look at the last layer of the hierarchy where each exemplar occupies its own cluster. However the hierarchical clustering offers the ability for a user to reduce the number of clusters, and therefore limit the redundancy, as may be required. We have coloured the exemplars in figure 3.6 according to clustering into 7 clusters.

## 3.5 Application to Breast Cancer Histology Images

### 3.5.1 Introduction

We show here a potential application of the seriation extension to AP to the problem of mitotic cell detection in breast cancer histopathology images [Khan et al., 2013]. The algorithmic framework presented by the authors of that work involved using pixel features to detect potential mitotic nuclear regions, then using a context aware post-processing (CAPP) to reduce the number of false detections, figure 3.7 shows some mitotic nuclear regions and some falsely detected regions. The full details of that work can be found in [Khan et al., 2013]. Here we investigate if shape

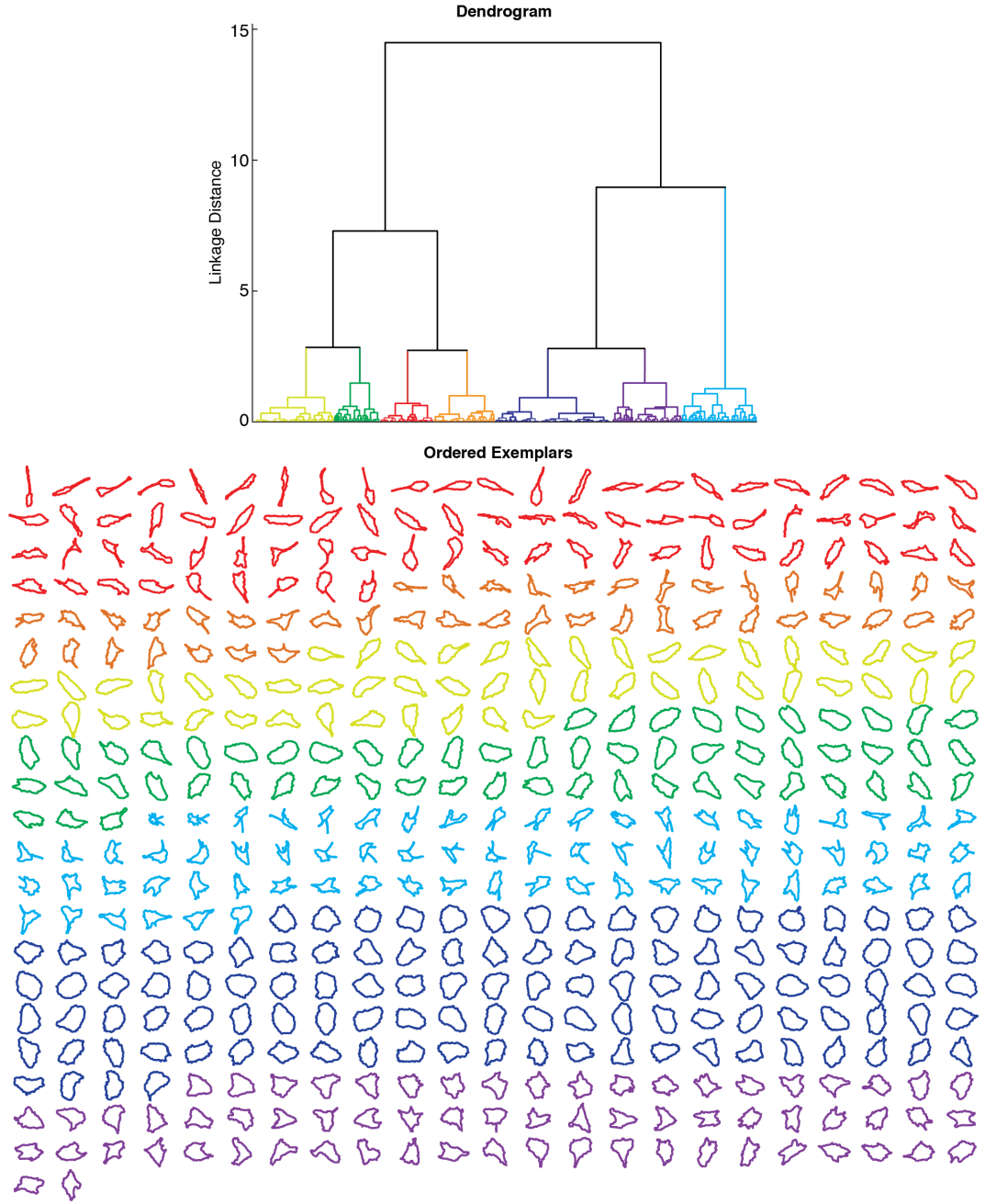


Figure 3.6: **Seriation Ordered Exemplars.** This figure shows the seriation re-ordering of exemplars generated by Affinity Propagation clustering on our set of 37818 cell shapes. The difference between this and figure 3.5 is that here, the exemplars have been reordered to best preserve shape similarity between near exemplars. At the top is a dendrogram showing the result of hierarchical clustering on the AP clusters, this is an intermediate step in reordering the exemplars. We show by colour the grouping of the exemplars into 7 clusters.

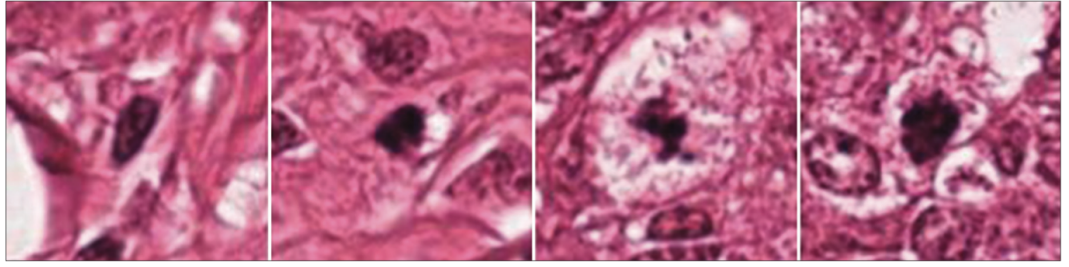


Figure 3.7: **Example Mitotic Cell Candidates.** Figure reproduced from [Khan et al., 2013]. Four examples of  $50 \times 50$  context patches, cropped around the bounding box of candidate mitotic nuclear regions (detected using the Gamma-Gaussian Mixture Model proposed in [Khan et al., 2013]). First 2 (from left) are false positives, last 2 are mitotic cells.

features could potentially help distinguish mitotic nuclear regions from other objects (eg, apoptotic nuclei, other non-mitotic cell nuclei, cell debris etc) that are falsely detected.

### 3.5.2 Applying the Extended Affinity Propagation to Histology Data

To attempt to find morphological features that distinguish mitotic nuclear regions from other structures in the tissue that have similar pixel features we applied the BAM based affinity propagation algorithm with the seriation extension to segmented candidate regions. This analysis was chosen in order to inspect the morphological distribution of the candidate regions and determine if there were features that could distinguish mitotic nuclear regions.

In figure 3.8 we show the 314 AP exemplars displayed from left to right, line by line, as ordered from the seriation extension. At the bottom of figure 3.8 we show 10 randomly selected exemplars again and 15 randomly selected members of each corresponding cluster, in order to provide evidence for the cluster integrity of the AP cluster model, i.e we claim members of each cluster show high shape similarity, or appear to come from the same morphology phenotype. However, as with the RPE dataset, we see redundancy in the AP cluster model, i.e. many exemplars represent similar shapes. As the shape space is continuous without hard cluster limits, it is impossible to reduce the redundancy without compromising the integrity of the clusters. The colours represent a coarse clustering using hierarchical clustering into 5 clusters. The value of the framework is that the seriation ordering reflects the idea that neighbouring clusters are not independent. We expect that any feature related to shape will be non-randomly structured when viewed against the distribution

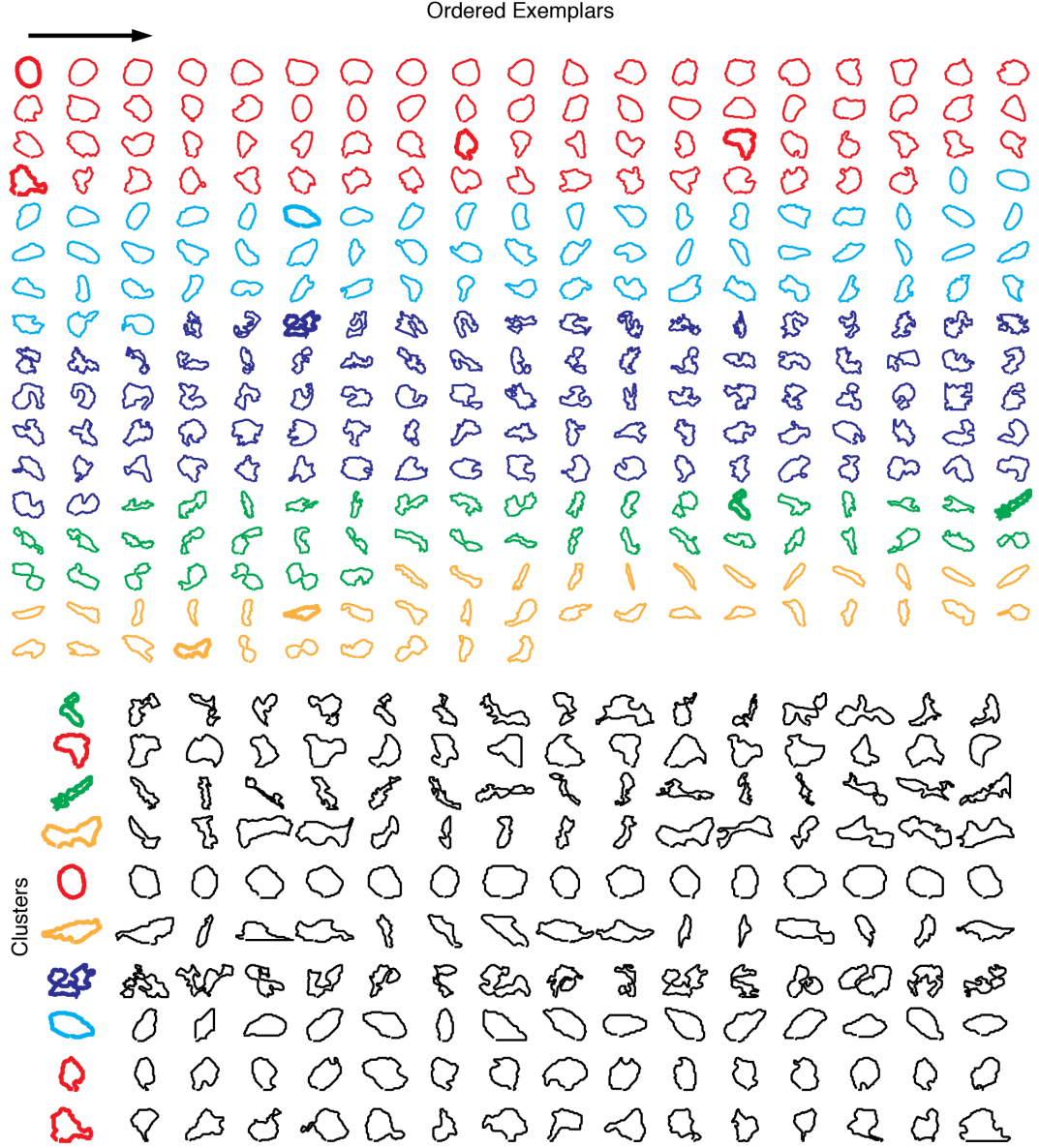


Figure 3.8: **Ordered Exemplars for Mitosis Dataset.** This figure shows the seriation reordering of exemplars generated by Affinity Propagation clustering on a set of potential mitotic nuclear regions. The exemplars have been reordered to best preserve shape similarity between near exemplars. Ten clusters were randomly chosen, and 15 elements were randomly chosen from each cluster to demonstrate the fidelity of the cluster assignment. We show by colour the grouping of the exemplars into 5 clusters.

created by the seriation framework.

In figure 3.9 we show a bar chart that illustrates the proportion of mitotic nuclear regions in each AP cluster. The order of the bars corresponds to the order generated by seriation. The colours of the bars corresponds to the coarse clustering into 5 clusters. If there were any way to identify mitotic regions based on shape, we would expect to see the mitotic populations in figure 3.9 to be grouped up in some way. However, as we can see the mitotic populations are arbitrarily distributed across our shape clusters, which suggests that shape gives no extra information about whether the detected regions are in fact mitotic.

### 3.6 Discussion

Section 3.4.3 was intended to assess the merits of our shape similarity measure, BAM, independently from the Diffusion Maps framework that we will be using later. We use Affinity Propagation [Frey and Dueck, 2007] with an extension inspired by the seriation algorithm [Wishart, 1999] to generate an ordered list of cell shape exemplars (see figure 3.6). We believe these exemplars are well structured by this process, showing similar phenotypes classed together and good separation of different shapes. Hence, we argue that BAM produces a sensible measure of the similarity of our cell shapes that at the very least reflects the human perception of shape similarity. In the next chapter we will look at the performance of BAM within the Diffusion Framework where we attempt to visualise the morphological features captured in the embedding process. The best vindication of BAM in this framework is clearly good performance in a suitable task, and this is what we attempt to do in chapter 5.

While the methodology presented in section 3.4.3 was primarily designed to be simply a validation of BAM, it can be viewed as a useful framework in it's own right. Firstly we argue that, with a huge dataset, hierarchical clustering performed after affinity propagation can be an effective method for reducing redundancy in the first cluster model. The method proposed uses the correlation between each of the exemplars' vectors of similarity scores and generates a hierarchy of cluster assignments, allowing a user to find an optimal cluster assignment to maximise inter- versus intra-cluster variation.

Another potentially useful part of the work in section 3.4.3 is the final order of the exemplars. This order should, in some optimised way, reflect shape similarity, with short sections of the list showing shapes with progressive values for some important shape feature(s). This means, if one were to suspect that the value of another vari-

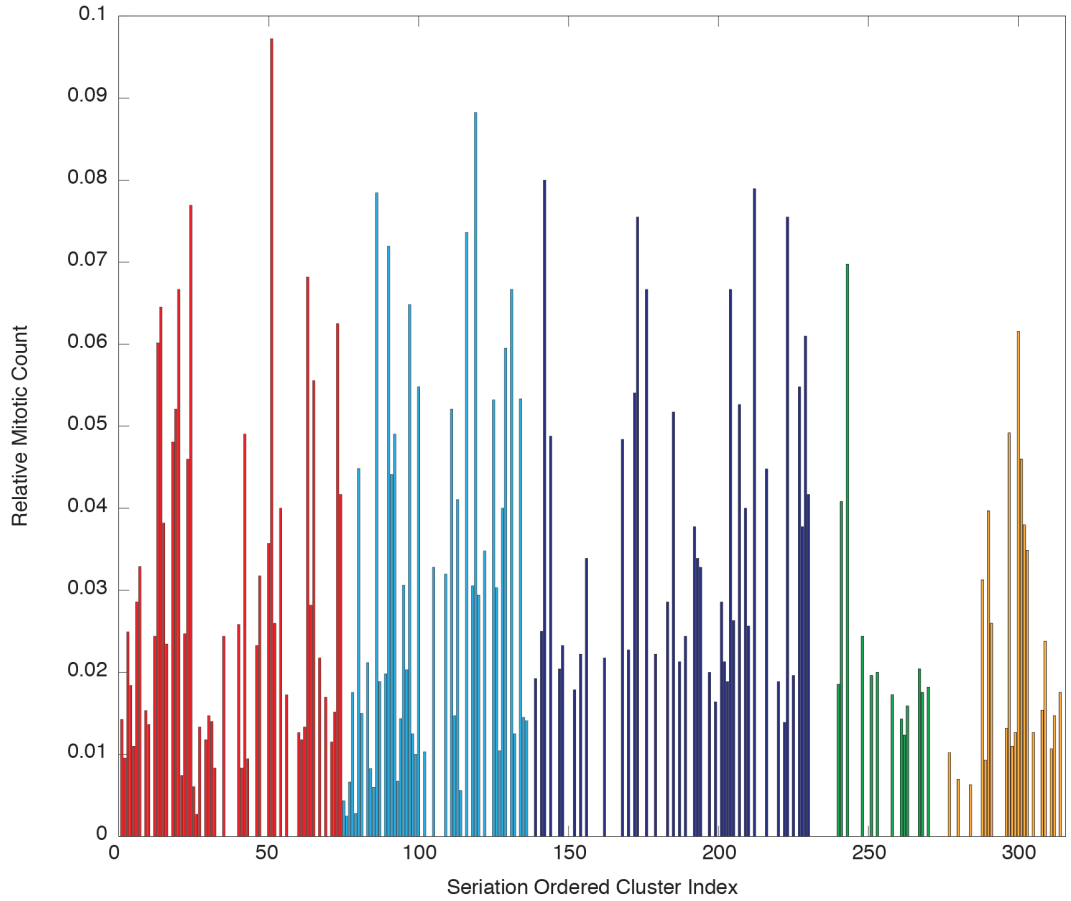


Figure 3.9: **Relative Mitotic Count in Ordered Clusters.** The proportion of true mitotic nuclear regions in each cluster of a set of potential regions. These clusters were generated using Affinity Propagation and reordered according to the seriation extension described in section 3.4.4. This cluster reordering is based on neighbouring clusters having similar shape features (see figure 3.8), hence if shape were any indication of whether a region is a mitotic nuclear region we would see some organisation in this figure.



able (such as the concentration of some molecule within each cell or its environment) is related to cell shape, one would hope to see structure when the value is plotted against this list. At least within the short sections.

## Chapter 4

# Morphological Phenotyping of Retinal Pigment Epithelial Cells

### 4.1 Visualising Shape Space

The result of applying our shape analysis framework (introduced in section 1.4) is a low dimensional embedding of our dataset of cell shapes where Euclidean distance reflects shape similarity as determined by our similarity measure, BAM, introduced in chapter 3. It is always challenging to assess the performance of a shape analysis algorithm since there is no ground truth when it comes to quantifying shape. In many investigations a first assessment is done visually and given that shape especially is often classified subjectively (see section 1.2.1) we wanted to develop some tools for visually inspecting the distribution created through Diffusion Maps.

What we expect from an embedding that preserves similarity is one where points are close together if they represent shapes that are similar, and points are far apart if the shapes are different. This should mean that if we look at local clusters of points in the embedding, the corresponding group of cell shapes should represent individual morphological phenotypes. Also if we look at points that lie in a straight line, such as one parallel to one of the axes, we should see (in the shapes) a progression of one or more shape features that are significant in the observed variability in the dataset.

#### 4.1.1 Shape Averaging

The simplest thing we can do to examine whether our embedding has achieved the properties described above is to look at the average shape of selected groups of cell

curves. Given a set of curves (scaled to standard path-length and interpolated to uniform parameterisation) and one chosen reference curve, we can use BAM to align each curve’s orientation and cyclic parameterisation with the reference curve (see section 3.2.2 and A). We can then simply find the average shape by computing the mean of all points of each index in all curves.

Figure 4.1 shows the result of partitioning our dataset according to intervals within each of the first two diffusion coordinates (separately). In red we can see the partitioning according to the first diffusion coordinate and the resulting average shapes. Clearly the most dominant shape feature that appears to vary in relation to this first axis has to do with aspect ratio; shapes with a very low first coordinate are as wide as they are long whereas shapes with a high first coordinate are much longer. This is consistent with the idea that cellular phenomena such as tail growth, are responsible for most of the morphological variation in the dataset. In green we see the result of partitioning based on the second diffusion coordinate.

#### 4.1.2 Extended Affinity Propagation to Visualise DM Embedding

In section 3.4.3 we look into using Affinity Propagation clustering on our large dataset of shapes and introduce an extension to it. This is a completely separate algorithm to the Diffusion Maps embedding algorithm, except for the use of the shape difference measure, BAM. The extension is necessary because the clustering results leave redundancy in the cluster model. We can make use of both the extended cluster model and the initial AP exemplars to evaluate the quality of the Diffusion framework. Firstly, the extension uses hierarchical clustering and can produce a coarse cluster assignment. We can apply this coarse clustering to the points generated through Diffusion Maps and examine the validity of this labelling to see if we have agreement between the two methods. This can be seen in the top plot of figure 4.2. The figure shows relatively good cluster integrity between the seven clusters, since the clusters stay mostly confined to their own region of the embedding. There is, however, some overlap, but we believe this to be inevitable considering that the data really is a continuum and does not naturally yield cluster boundaries.

The second plot in figure 4.2, looks at where the exemplars lie in the Diffusion embedding which allows us to visually check for good separation according to shape, and look for clues as to the important shape features controlling the newly generated Diffusion coordinates. Firstly, let us look at the separation according to shape; we can see (as before, in 3.6) that each cluster seems, to a certain degree, to represent a different morphological phenotype, as we see distinguishing properties in each. But what we see here is more of an idea of the continuity of the data. Firstly, within

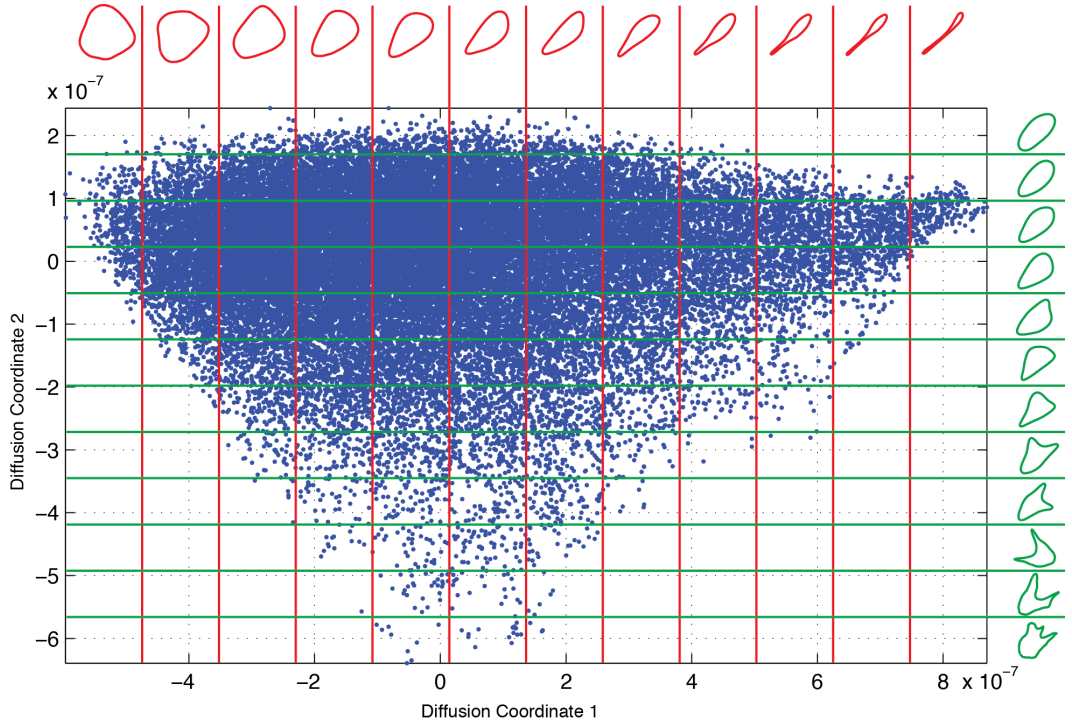


Figure 4.1: **Axis phenotypes.** This figure provides a coarse look at the shape features captured by each of the first two Diffusion coordinates. The underlying plot shows 37818 points each representing a cell's curve, arranged in such a way as to reflect shape similarity through Euclidean distance (section 2.3 outlines the Diffusion Maps algorithm, chapter 3 discusses the specifications made for this dataset). The figure also displays the average shape of subsets of shapes created by partitioning according to the first two Diffusion coordinates (separately). Red shows the partitioning according to the first coordinate, green shows the partitioning according to the second coordinate.

each cluster, in that subtle differences can be seen as you look across each cluster. Secondly, on the boundaries between clusters, we really can appreciate here that hard boundaries are not appropriate since clear similarities can be seen by many pairs of shapes that lie close to each other near the boundaries. Thirdly, in the arrangement of the clusters, we can see overarching progression of different features. The third point above is what we can use to help us interpret the features captured by the embedding. For example, as you look from (dark) blue to green to yellow and to red, there is clearly a shape progression relating to the same features we noticed in figure 4.1, that of round shapes versus long and thin shapes. Also, if we look at the yellow and orange clusters, these are both neighbours of the red cluster. Appropriately, we can say that the yellow and orange clusters share similarities

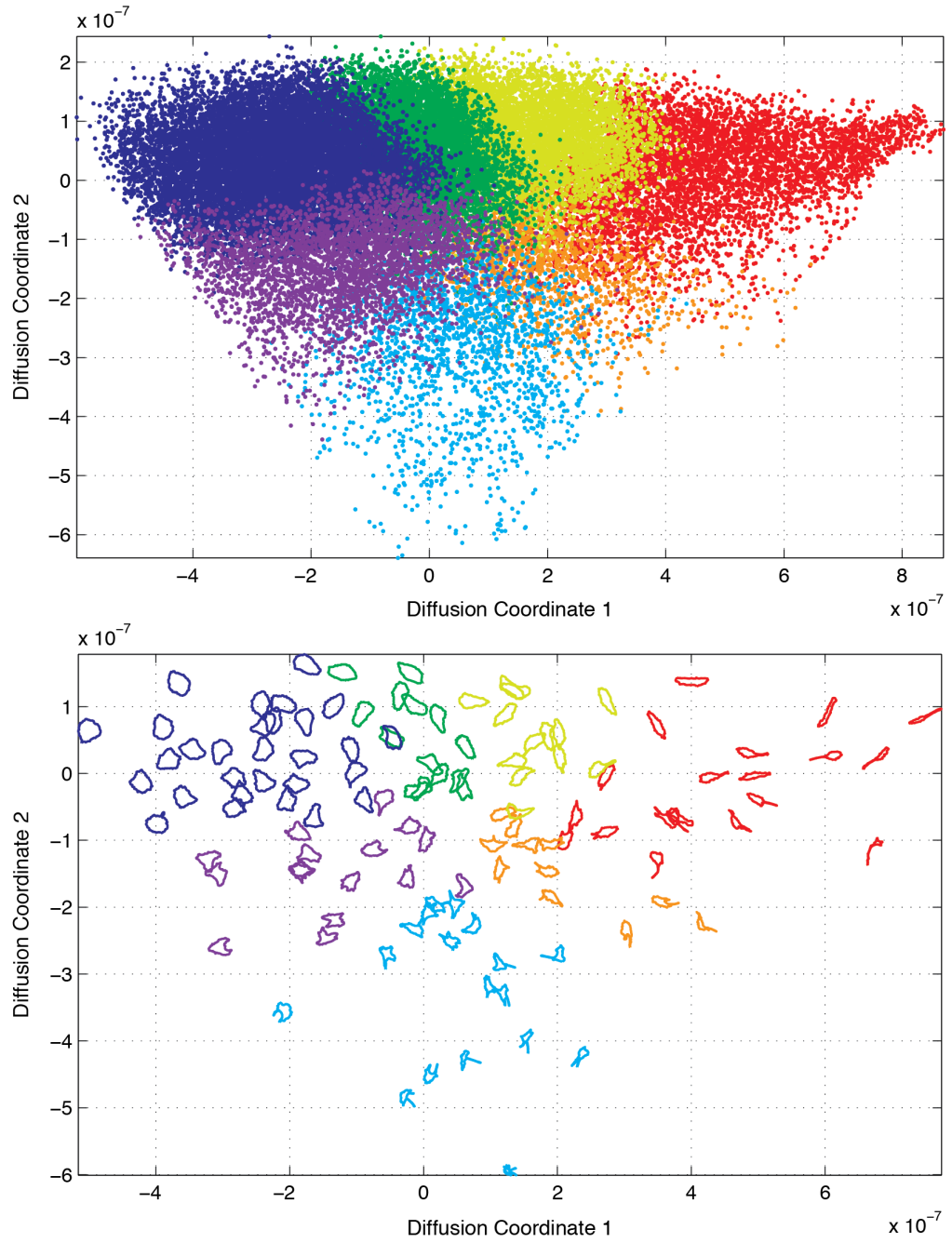


Figure 4.2: **Affinity Propagation Clusters over DM Embedding.** The top figure depicts the correspondence between the Diffusion Maps embedding (the  $xy$ -position) and the cluster assignment by Affinity Propagation (the colour). The bottom figure displays some of the AP exemplars overlaid onto the corresponding DM embedded points.

with the red cluster, but have differences which distinguish them, namely differing degrees of symmetry and convexity in the shapes. Similar differences can be seen in the green and purple clusters, and both are similar to the (dark) blue cluster. The light blue (or cyan) is clearly identified by the Diffusion framework as being removed from the others in the second coordinate. There is also a clear distinction in the shapes within this cluster, with many having very irregular shapes and numerous processes. In the next section we further investigate the features represented by our newly generated Diffusion coordinates.

## 4.2 Shape Feature Correlation

### 4.2.1 Scalar Shape Features

This section investigates the extent to which the Diffusion Maps shape representation correlates with simple shape features. We weren't too selective of the features we examined since we wanted to be very inclusive. Many of the features were computed using MATLAB's `regionprops` toolbox. To compute shape features of a given cell curve we construct a binary image from that curve and its bounding box. We refer to the part of the image inside the curve as the 'region' and define the features as follows:

- **Area:** The actual number of pixels in the region.
- **Major Axis Length:** Scalar specifying the length (in pixels) of the major axis of the ellipse that has the same normalised second central moments as the region.
- **Minor Axis Length:** Scalar specifying the length (in pixels) of the minor axis of the ellipse that has the same normalised second central moments as the region.
- **Eccentricity:** Scalar that specifies the eccentricity of the ellipse that has the same second-moments as the region. The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length. The value is between 0 and 1. (0 and 1 are degenerate cases; an ellipse whose eccentricity is 0 is actually a circle, while an ellipse whose eccentricity is 1 is a line segment.)
- **Orientation:** The angle (in degrees ranging from -90 to 90 degrees) between the x-axis and the major axis of the ellipse that has the same second-moments as the region.

- **Convex Area:** Scalar that specifies the number of pixels in the convex hull.
- **Solidity:** Scalar specifying the proportion of the pixels in the convex hull that are also in the region. Computed as  $\text{Area}/\text{Convex Area}$ .
- **Extent:** Scalar that specifies the ratio of pixels in the region to pixels in the total bounding box. Computed as the total area divided by the area of the bounding box.
- **Perimeter:** The distance around the boundary of the region. Regionprops computes the perimeter by calculating the distance between each adjoining pair of pixels around the border of the region.
- **Circularity:** Measured by computing  $P/(2\sqrt{\pi A})$  where  $P$  is the perimeter and  $A$  is the area.
- **Symmetry:** Scalar specifying the ratio of pixels bounded by both the cell curve and its reflection in its major axis to the number of pixels bounded by the cell curve.
- **Max distance from centre:** The maximum distance between the centre of mass (of the region bounded by the cell's boundary) and any point on the boundary of the cell.
- **Min distance from centre:** The minimum distance between the centre of mass (of the region bounded by the cell's boundary) and any point on the boundary of the cell.
- **Min/max centre distance ratio:** The ratio of the minimum to the maximum distances between the centre of mass and any point on the boundary of the cell.
- **Irregularity:** Also described as non-circularity, irregularity is computed as 
$$\frac{1 + \sqrt{\pi} \max_i \sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2}}{\sqrt{\text{Area}}} - 1.$$
- **Irregularity 2:** This is the irregularity of the negative space of the cell region within the bounding box.

Simple Shape Feature	Correlation with Diffusion Coordinates				
	D.C.1	D.C.2	D.C.3	D.C.4	D.C.5
Area	0.1882	0.1888	0.0869	0.0141	0.0063
Major Axis Length	0.7949	-0.0714	0.0411	0.0835	-0.0039
Minor Axis Length	-0.3844	-0.5346	-0.1235	-0.0735	0.0130
Eccentricity	0.8303	0.2197	0.0740	0.1650	-0.0369
Orientation	-0.0035	-0.0261	-0.0091	0.0111	0.0052
Convex Area	0.4258	-0.4339	-0.0096	0.0349	0.0079
Solidity	-0.6427	0.6315	0.1268	-0.0382	0.0024
Extent	-0.7858	0.4063	0.1378	0.0071	0.0016
Perimeter	0.6643	-0.4763	0.0498	0.0057	0.0115
Circularity	0.7664	-0.4899	-0.0040	0.0181	-0.0559
Symmetry	-0.4144	0.6222	0.2296	-0.1220	0.0090
Max distance from centre	0.8049	-0.2267	-0.0853	-0.0860	0.0063
Min distance from centre	-0.5287	0.0220	0.1607	-0.2799	0.0338
Min/max centre distance ratio	-0.9053	0.2401	0.1629	-0.0979	0.0343
Irregularity	0.8530	-0.1703	-0.1468	-0.0923	0.0033
Irregularity2	-0.5387	0.3960	0.0363	-0.0811	-0.0157

Table 4.1: This table shows the correlation of simple shape features with the Diffusion Maps representation of our RPE1 dataset. The shape features are described in section 4.2.1. The Diffusion embedding is described in section 2.3

#### 4.2.2 Shape feature correlation

Table 4.1 displays the correlation between each of our Diffusion coordinates with the shape features described in section 4.2.1, where correlation is computed as

$$r_{x,y} = \frac{\sum_{t=1}^N (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^N (x_t - \bar{x})^2 \cdot \sum_{t=1}^N (y_t - \bar{y})^2}}, \quad (4.1)$$

where  $N=37818$ ,  $x_t, y_t$  represent the relevant Diffusion coordinate or shape feature value for curve  $t$  and  $\bar{x}, \bar{y}$  represent the respective means. From table 4.1 we see that many of these features are strongly correlated with our new Diffusion coordinates. The features that correlate most strongly with the first Diffusion coordinate include the ratio between the maximum and minimum distances from the border to the centre of gravity, as well as irregularity and eccentricity. These all seem to have a similar theme about the cell being longer than it is wide. This corresponds with what we had noticed about the distribution in inspection in section 4.1. Bearing in mind that the Diffusion Maps algorithm is designed to find the most important sources of variability within the dataset, this result closely matches our biological expectation, which was that a cell being polarised (having a prominent front end



and tail) or not is the most significant morphological feature.

The features that correlate most strongly with the second diffusion parameter are solidity and symmetry. Remember that solidity is computed as  $\frac{\text{Area}}{\text{Convex Hull Area}}$  and symmetry is computed as  $\frac{\text{Reflected Intersection Area}}{\text{Area}}$ . Both of these statistics reflect (among other irregularities) the presence of side processes or other structures that give the cell a branched shape. This again fits both our earlier investigation (section 4.1) and our biological expectation as being an important feature in the cells' behaviour.

One interesting utility of this table is in investigating the success of the dimension reduction process. We can actually use the information in this table to determine how many new dimensions it is appropriate to use. For example, we can look at top 5 features that correlate with the 2nd Diffusion Coordinate, namely, Solidity, Symmetry, Minor Axis Length, Circularity and Perimeter (Note, here, we are ordering according to the absolute value of the correlation coefficient). The position of these features in terms of correlation with the first coordinate are respectively 9th, 13th, 14th, 7th and 8th. The fact that these positions are relatively low implies that the second Diffusion Coordinate can be interpreted as representing very different shape information to the first Diffusion Coordinate. Whereas, we can look at the top 5 features correlated with the third Diffusion Coordinate and their highest positions in respect to either of the first two Diffusion Coordinates; i.e Symmetry is 2nd (for DC2), Min/max centre distance ratio is 1st (for DC1), Minimum centre distance is 11th (for DC1), Irregularity is 2nd (for DC1) and Extent is 6th (DC1). With the exception of Minimum centre distance, these are all quite high positions which suggests that the shape information captured by the third dimension is largely already captured by the first two dimensions. The minimum centre distance (the smallest distance between a curve and its centre) is a feature quite unlike the other features, and perhaps the third dimension has successfully separated this information from the rest. However at its low correlation value it is likely that this feature does not account for much variation within the dataset. Similarly, the top 5 features correlating with the fourth Diffusion Coordinate achieve positions 3, 3, 1, 1 and 2 amongst the first three Diffusion Coordinates and the top 5 features correlating with the fifth Diffusion Coordinate achieve positions 4, 2, 1, 1 and 8 amongst the first four Diffusion Coordinates. These are almost all low positions, suggesting the information in these coordinates is already captured.

It is reassuring to see that the orientation of the cells has very low correlation with all of the Diffusion coordinates since our shape similarity measure was designed to be invariant to orientation. What is surprising is that features relating to scale

(Perimeter and Area) correlate with the Diffusion coordinates to the extent that they do. Since (by necessity rather than design) the cell shapes were analysed with a standardised perimeter. What this implies is that the perimeter is closely related to another shape feature (or perhaps many), from which it can be inferred. In section 4.2.3 we attempt to visualise the distribution of these features over the top two Diffusion coordinates.

### 4.2.3 Features distributed over the Diffusion Maps embedding

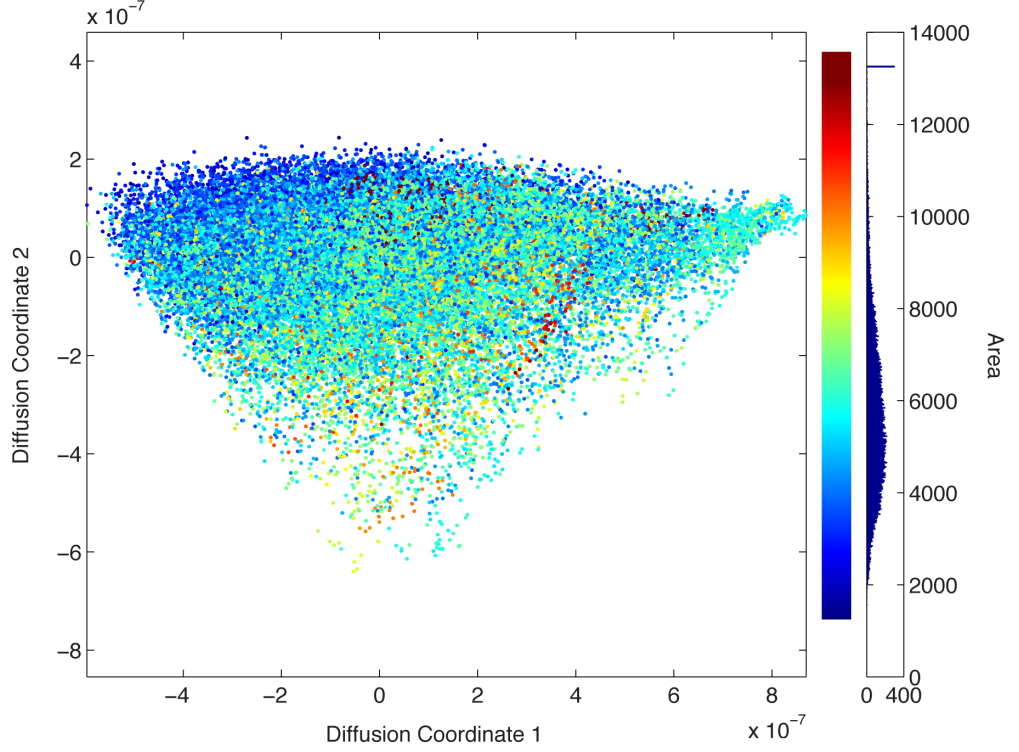


Figure 4.3: **Area distributed over DM embedding.** This figure looks at the distribution of cell shape *area* over the first two Diffusion Coordinates. Each point of the plot represents one cell shape, its position corresponds to its embedding through the Diffusion Maps process, its colour corresponds to its relative *area*. The colour map was generated to display the *area* range, truncated at 3.5 standard deviations either side of the median. Shown in the figure is a histogram showing the *area* distribution over the dataset and a colour bar that corresponds with the *area* value bins. The definition and computation of *area* are explained in section 4.2.1.

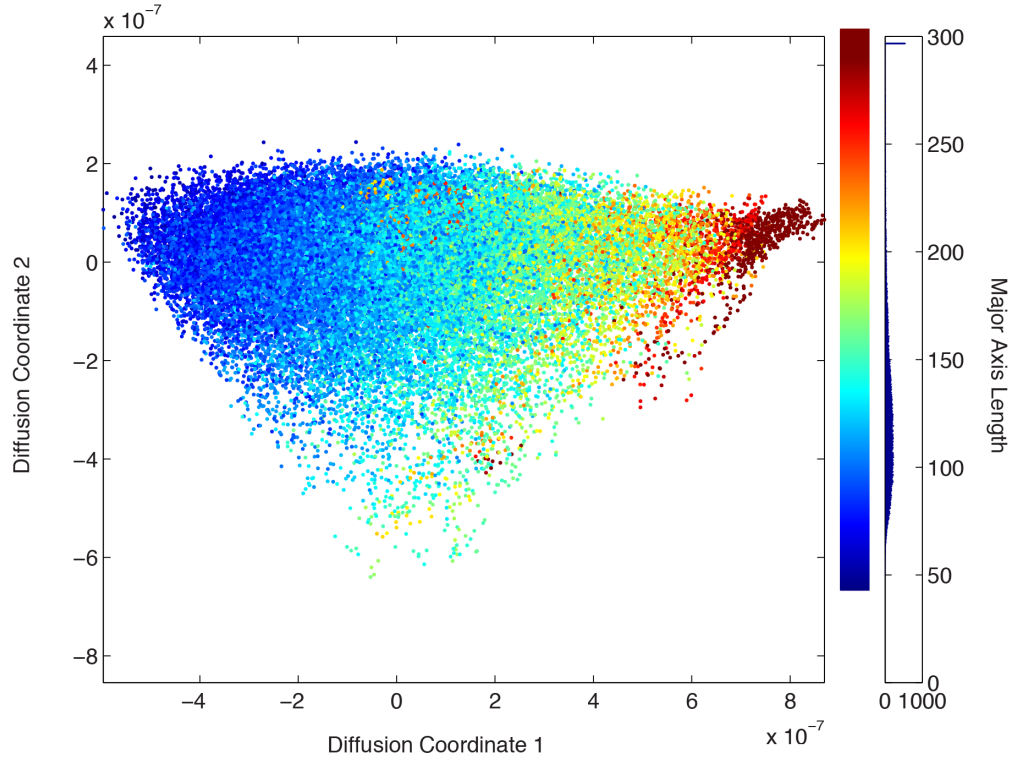


Figure 4.4: **Major Axis Length distributed over DM embedding.** This figure looks at the distribution of cell shape *major axis length* over the first two Diffusion Coordinates. Each point of the plot represents one cell shape, its position corresponds to its embedding through the Diffusion Maps process, its colour corresponds to its relative *major axis length*. The colour map was generated to display the *major axis length* range, truncated at 3.5 standard deviations either side of the median. Shown in the figure is a histogram showing the *major axis length* distribution over the dataset and a colour bar that corresponds with the *major axis length* value bins. The definition and computation of *major axis length* are explained in section 4.2.1.

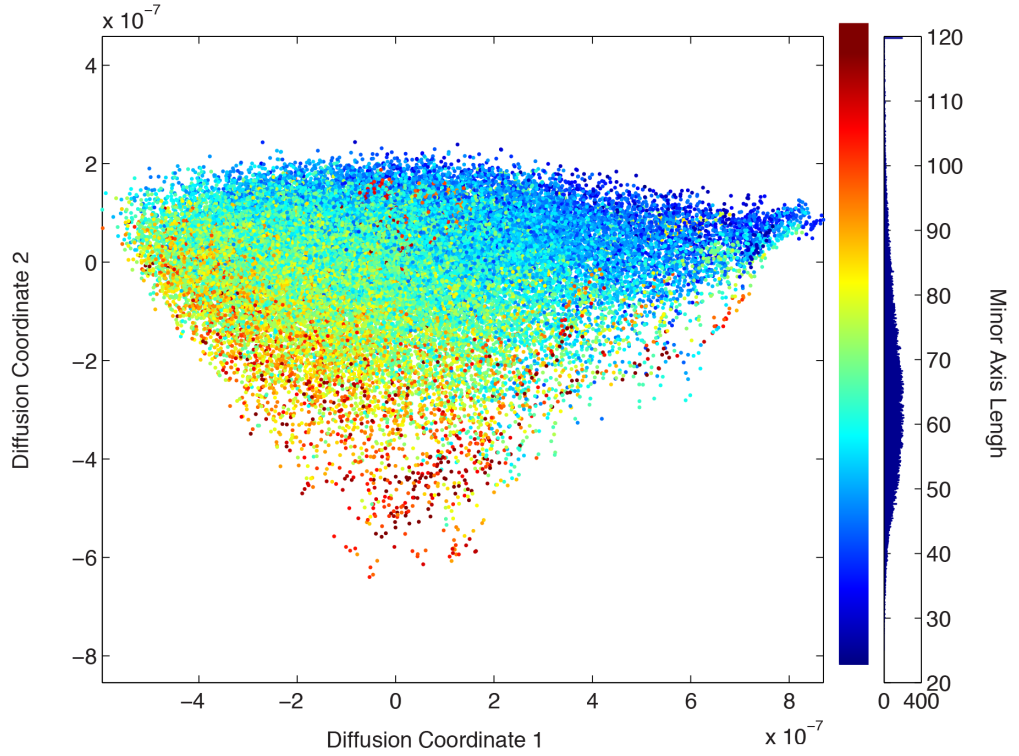


Figure 4.5: **Minor Axis Length distributed over DM embedding.** This figure looks at the distribution of cell shape *minor axis length* over the first two Diffusion Coordinates. Each point of the plot represents one cell shape, its position corresponds to its embedding through the Diffusion Maps process, its colour corresponds to its relative *minor axis length*. The colour map was generated to display the *minor axis length* range, truncated at 3.5 standard deviations either side of the median. Shown in the figure is a histogram showing the *minor axis length* distribution over the dataset and a colour bar that corresponds with the *minor axis length* value bins. The definition and computation of *minor axis length* are explained in section 4.2.1.

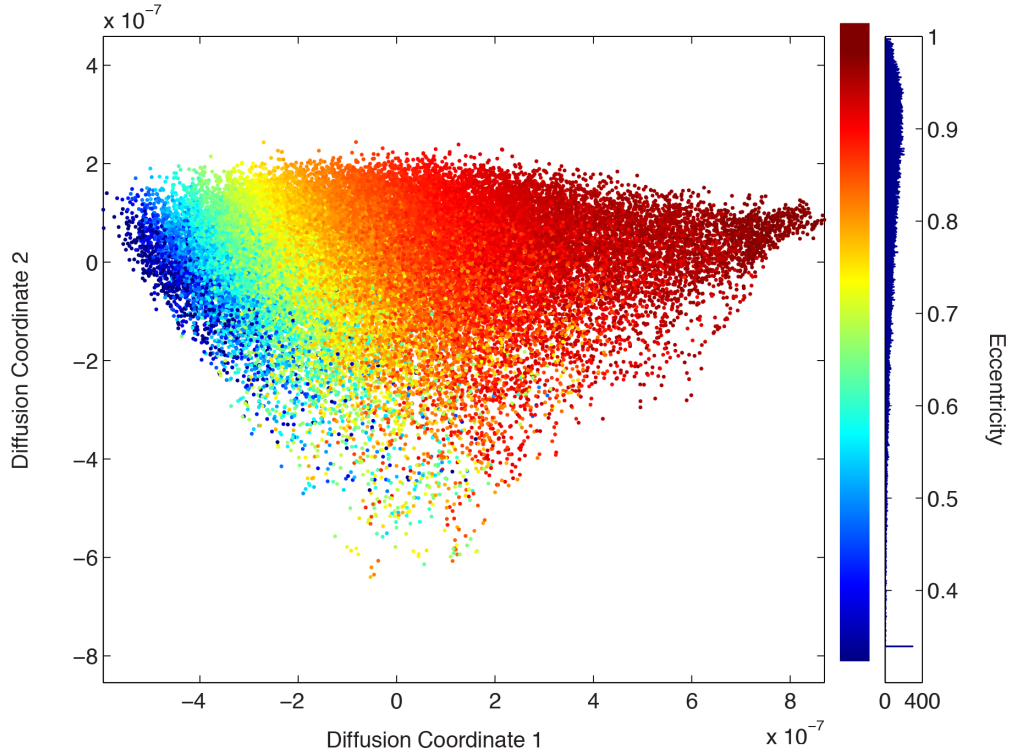


Figure 4.6: **Eccentricity distributed over DM embedding.** This figure looks at the distribution of cell shape *eccentricity* over the first two Diffusion Coordinates. Each point of the plot represents one cell shape, its position corresponds to its embedding through the Diffusion Maps process, its colour corresponds to its relative *eccentricity*. The colour map was generated to display the *eccentricity* range, truncated at 3.5 standard deviations either side of the median. Shown in the figure is a histogram showing the *eccentricity* distribution over the dataset and a colour bar that corresponds with the *eccentricity* value bins. The definition and computation of *eccentricity* are explained in section 4.2.1.

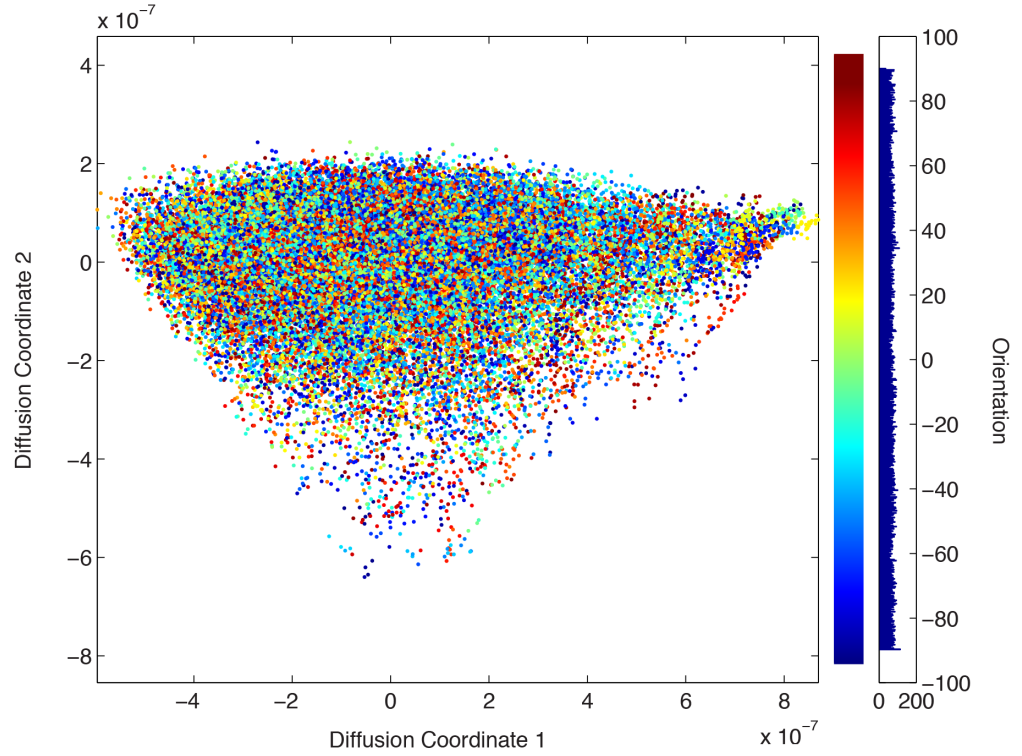


Figure 4.7: **Orientation distributed over DM embedding.** This figure looks at the distribution of cell shape *orientation* over the first two Diffusion Coordinates. Each point of the plot represents one cell shape, its position corresponds to its embedding through the Diffusion Maps process, its colour corresponds to its relative *orientation*. The colour map was generated to display the *orientation* range, truncated at 3.5 standard deviations either side of the median. Shown in the figure is a histogram showing the *orientation* distribution over the dataset and a colour bar that corresponds with the *orientation* value bins. The definition and computation of *orientation* are explained in section 4.2.1.

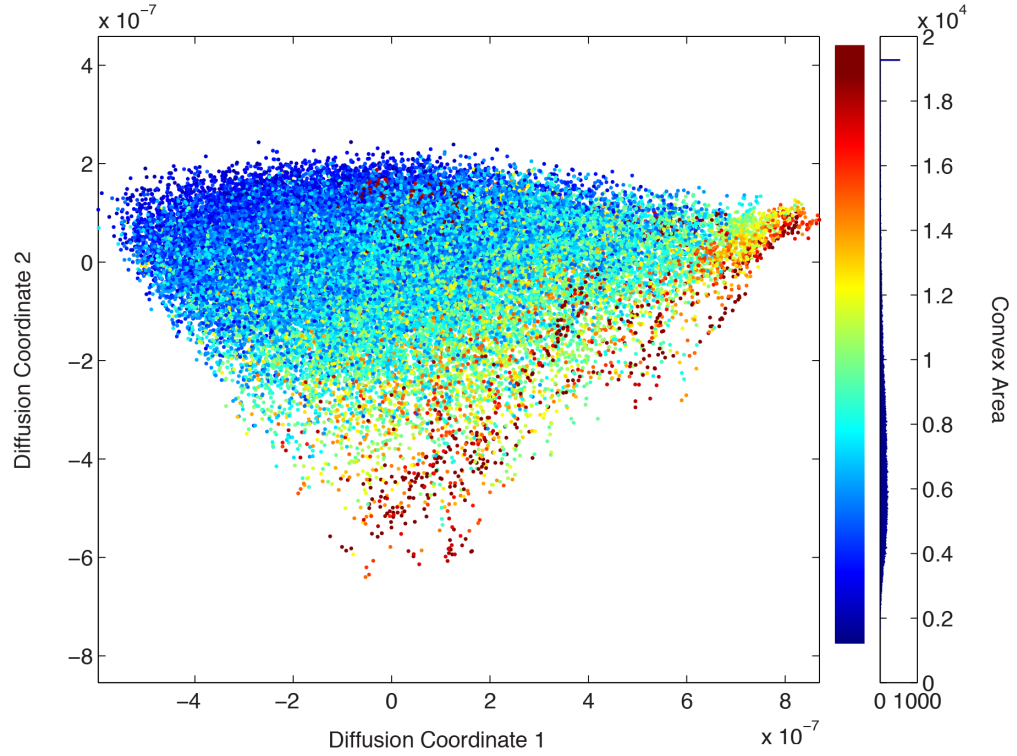


Figure 4.8: **Convex Area distributed over DM embedding.** This figure looks at the distribution of cell shape *convex area* over the first two Diffusion Coordinates. Each point of the plot represents one cell shape, its position corresponds to its embedding through the Diffusion Maps process, its colour corresponds to its relative *convex area*. The colour map was generated to display the *convex area* range, truncated at 3.5 standard deviations either side of the median. Shown in the figure is a histogram showing the *convex area* distribution over the dataset and a colour bar that corresponds with the *convex area* value bins. The definition and computation of *convex area* are explained in section 4.2.1.



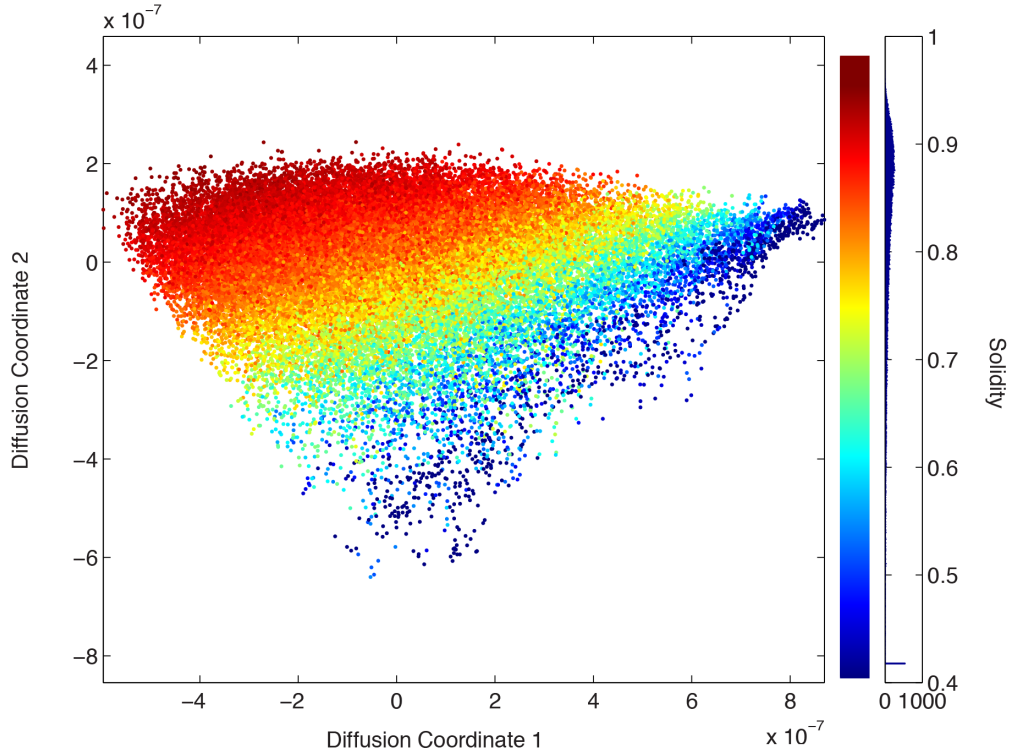


Figure 4.9: **Solidity distributed over DM embedding.** This figure looks at the distribution of cell shape *solidity* over the first two Diffusion Coordinates. Each point of the plot represents one cell shape, its position corresponds to its embedding through the Diffusion Maps process, its colour corresponds to its relative *solidity*. The colour map was generated to display the *solidity* range, truncated at 3.5 standard deviations either side of the median. Shown in the figure is a histogram showing the *solidity* distribution over the dataset and a colour bar that corresponds with the *solidity* value bins. The definition and computation of *solidity* are explained in section 4.2.1.

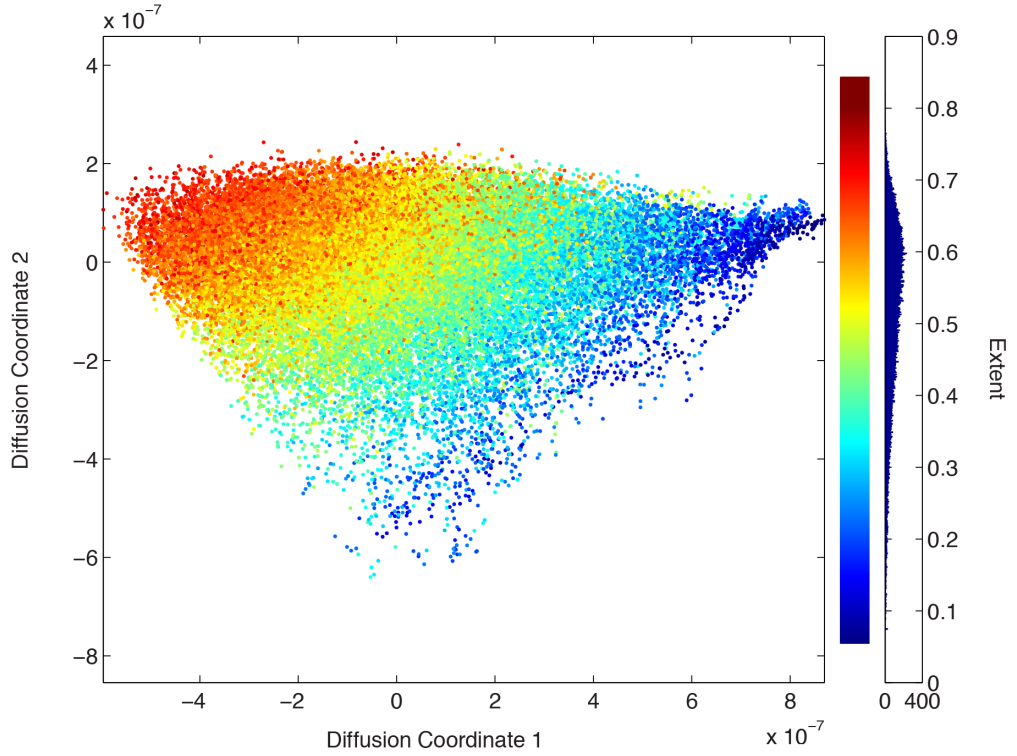


Figure 4.10: **Extent distributed over DM embedding.** This figure looks at the distribution of cell shape *extent* over the first two Diffusion Coordinates. Each point of the plot represents one cell shape, its position corresponds to its embedding through the Diffusion Maps process, its colour corresponds to its relative *extent*. The colour map was generated to display the *extent* range, truncated at 3.5 standard deviations either side of the median. Shown in the figure is a histogram showing the *extent* distribution over the dataset and a colour bar that corresponds with the *extent* value bins. The definition and computation of *extent* are explained in section 4.2.1.

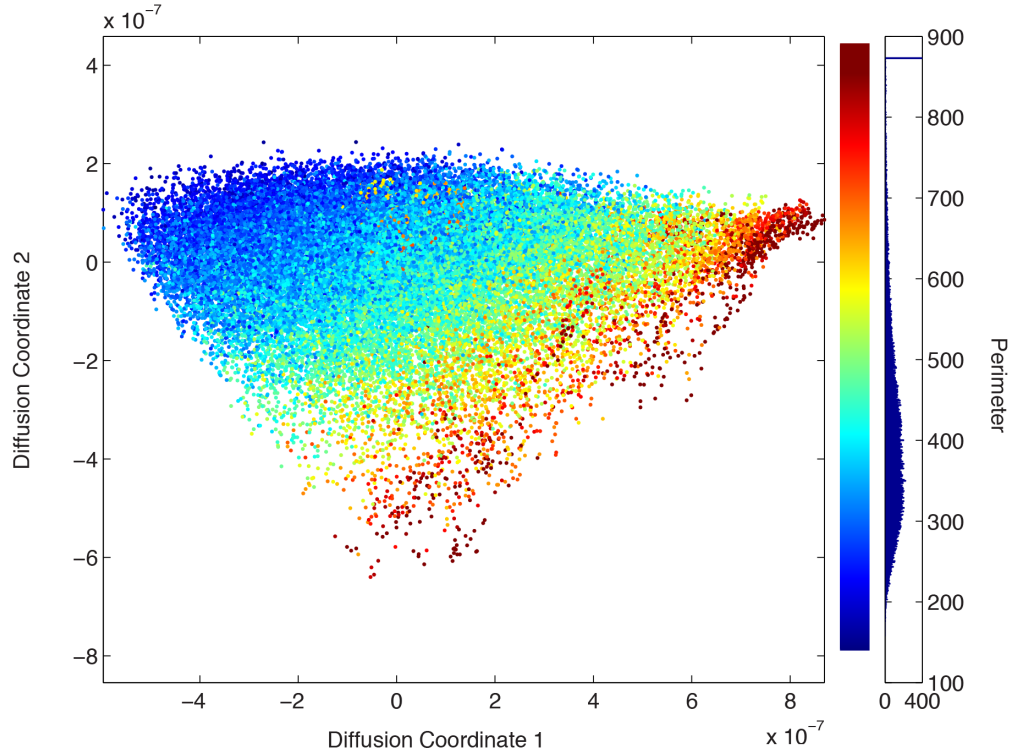


Figure 4.11: **Perimeter distributed over DM embedding.** This figure looks at the distribution of cell shape *perimeter* over the first two Diffusion Coordinates. Each point of the plot represents one cell shape, its position corresponds to its embedding through the Diffusion Maps process, its colour corresponds to its relative *perimeter*. The colour map was generated to display the *perimeter* range, truncated at 3.5 standard deviations either side of the median. Shown in the figure is a histogram showing the *perimeter* distribution over the dataset and a colour bar that corresponds with the *perimeter* value bins. The definition and computation of *perimeter* are explained in section 4.2.1.

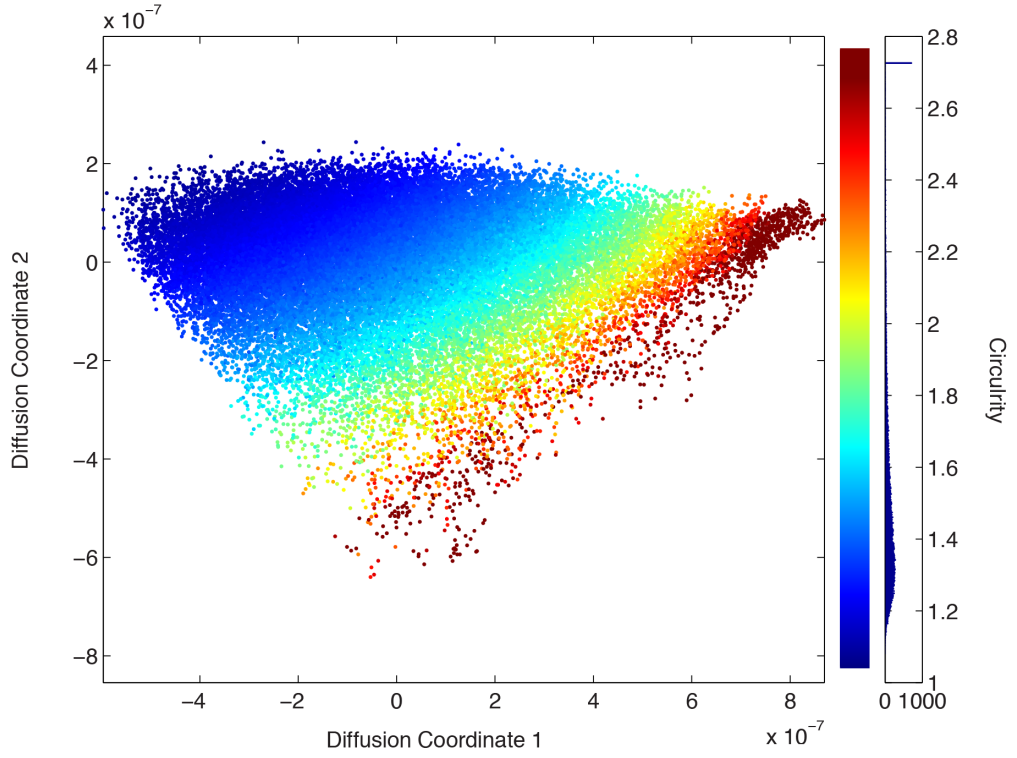


Figure 4.12: **Circularity distributed over DM embedding.** This figure looks at the distribution of cell shape *circularity* over the first two Diffusion Coordinates. Each point of the plot represents one cell shape, its position corresponds to its embedding through the Diffusion Maps process, its colour corresponds to its relative *circularity*. The colour map was generated to display the *circularity* range, truncated at 3.5 standard deviations either side of the median. Shown in the figure is a histogram showing the *circularity* distribution over the dataset and a colour bar that corresponds with the *circularity* value bins. The definition and computation of *circularity* are explained in section 4.2.1.

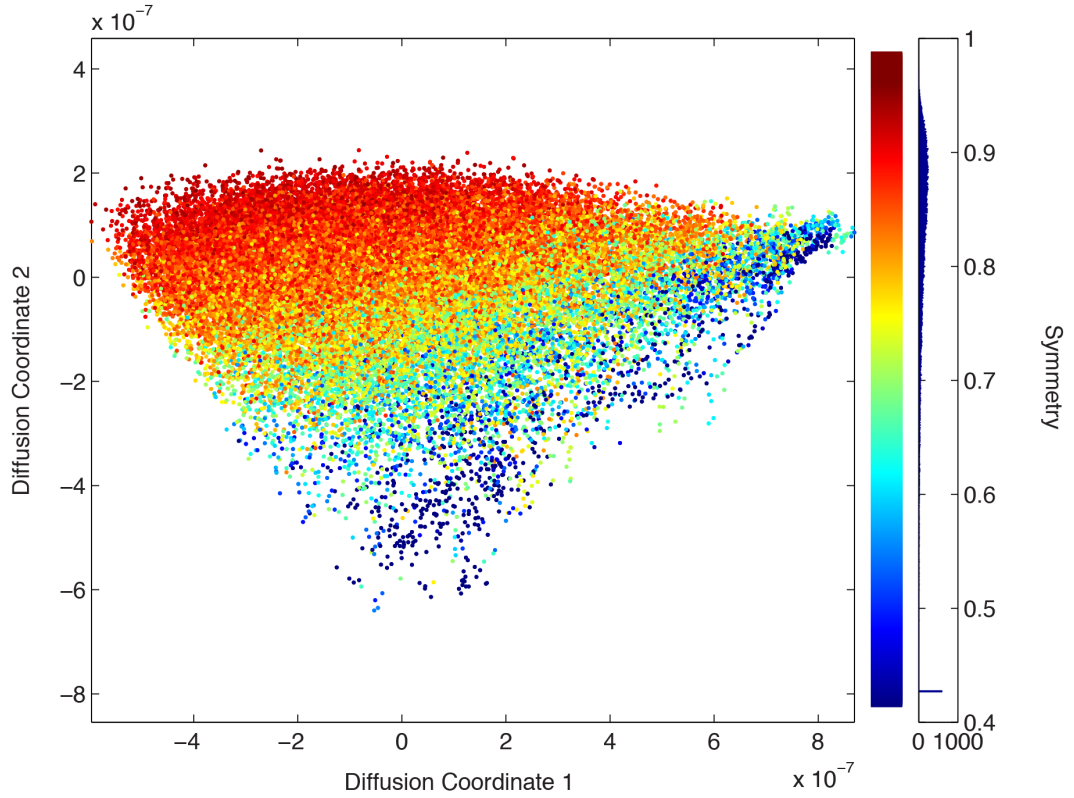


Figure 4.13: **Symmetry distributed over DM embedding.** This figure looks at the distribution of cell shape *symmetry* over the first two Diffusion Coordinates. Each point of the plot represents one cell shape, its position corresponds to its embedding through the Diffusion Maps process, its colour corresponds to its relative *symmetry*. The colour map was generated to display the *symmetry* range, truncated at 3.5 standard deviations either side of the median. Shown in the figure is a histogram showing the *symmetry* distribution over the dataset and a colour bar that corresponds with the *symmetry* value bins. The definition and computation of *symmetry* are explained in section 4.2.1.

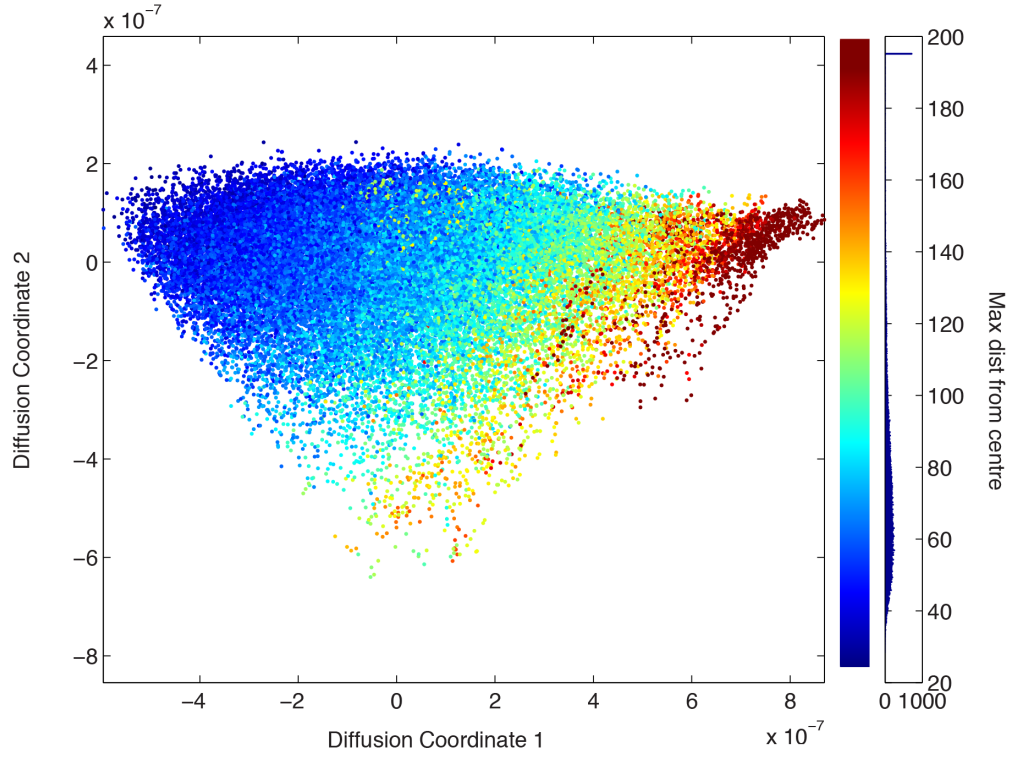


Figure 4.14: **Max distance from centre distributed over DM embedding.** This figure looks at the distribution of cell shape *max distance from centre* over the first two Diffusion Coordinates. Each point of the plot represents one cell shape, its position corresponds to its embedding through the Diffusion Maps process, its colour corresponds to its relative *max distance from centre*. The colour map was generated to display the *max distance from centre* range, truncated at 3.5 standard deviations either side of the median. Shown in the figure is a histogram showing the *max distance from centre* distribution over the dataset and a colour bar that corresponds with the *max distance from centre* value bins. The definition and computation of *max distance from centre* are explained in section 4.2.1.

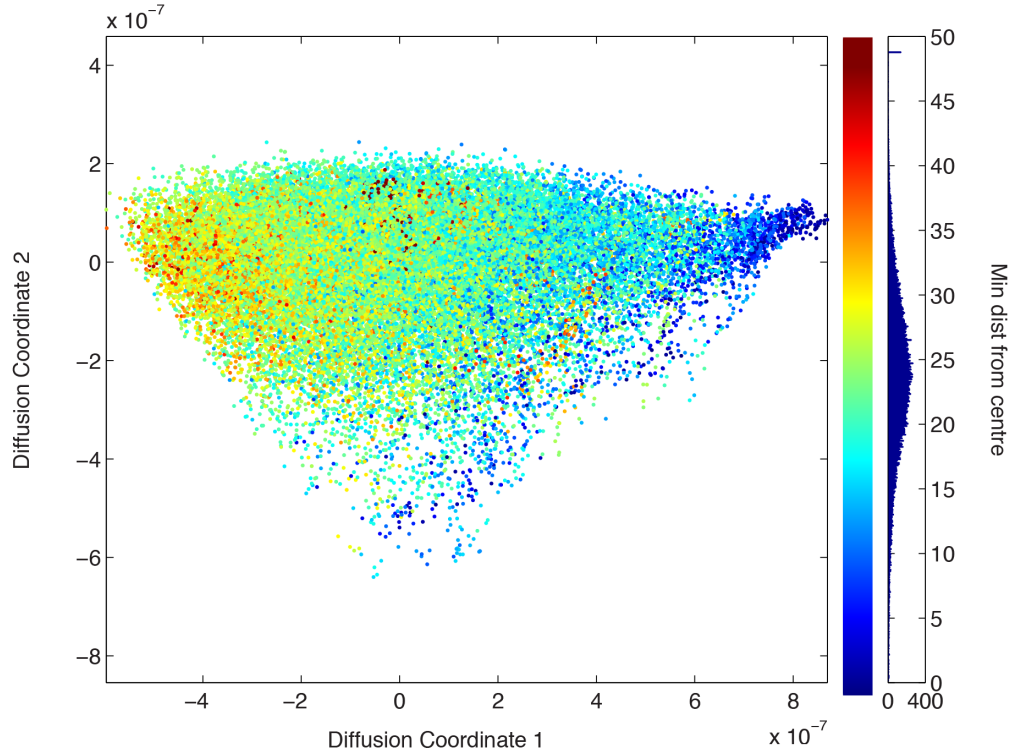


Figure 4.15: **Min distance from centre distributed over DM embedding.** This figure looks at the distribution of cell shape *min distance from centre* over the first two Diffusion Coordinates. Each point of the plot represents one cell shape, its position corresponds to its embedding through the Diffusion Maps process, its colour corresponds to its relative *min distance from centre*. The colour map was generated to display the *min distance from centre* range, truncated at 3.5 standard deviations either side of the median. Shown in the figure is a histogram showing the *min distance from centre* distribution over the dataset and a colour bar that corresponds with the *min distance from centre* value bins. The definition and computation of *min distance from centre* are explained in section 4.2.1.



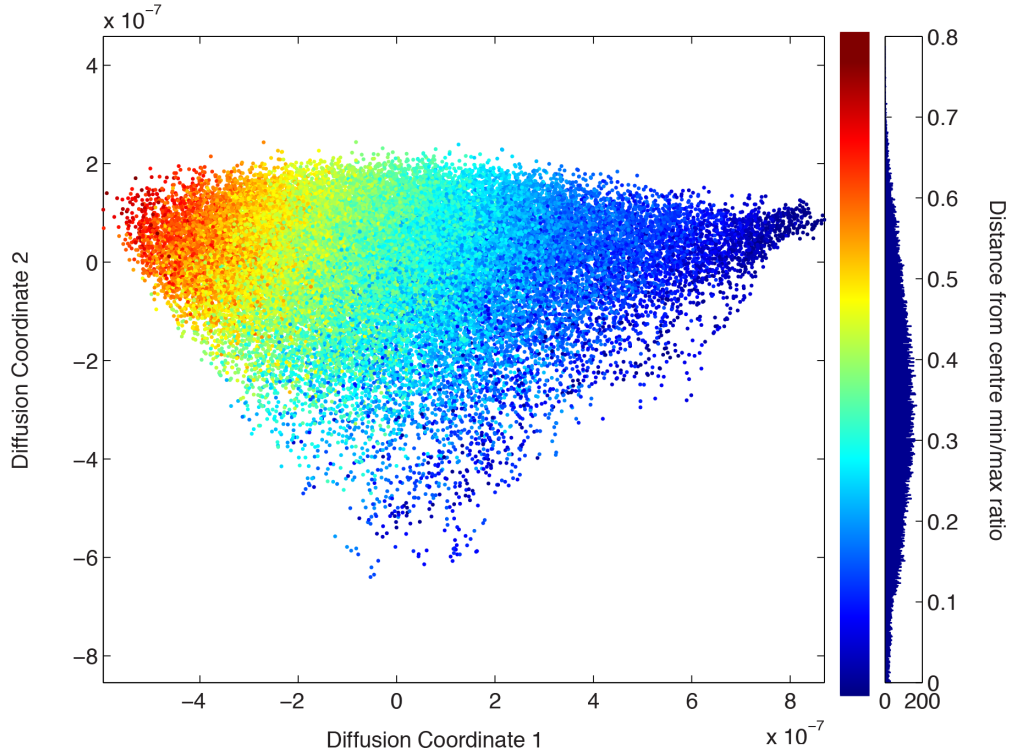


Figure 4.16: **Min/max centre distance ratio distributed over DM embedding.** This figure looks at the distribution of cell shape *min/max centre distance ratio* over the first two Diffusion Coordinates. Each point of the plot represents one cell shape, its position corresponds to its embedding through the Diffusion Maps process, its colour corresponds to its relative *min/max centre distance ratio*. The colour map was generated to display the *min/max centre distance ratio* range, truncated at 3.5 standard deviations either side of the median. Shown in the figure is a histogram showing the *min/max centre distance ratio* distribution over the dataset and a colour bar that corresponds with the *min/max centre distance ratio* value bins. The definition and computation of *min/max centre distance ratio* are explained in section 4.2.1.



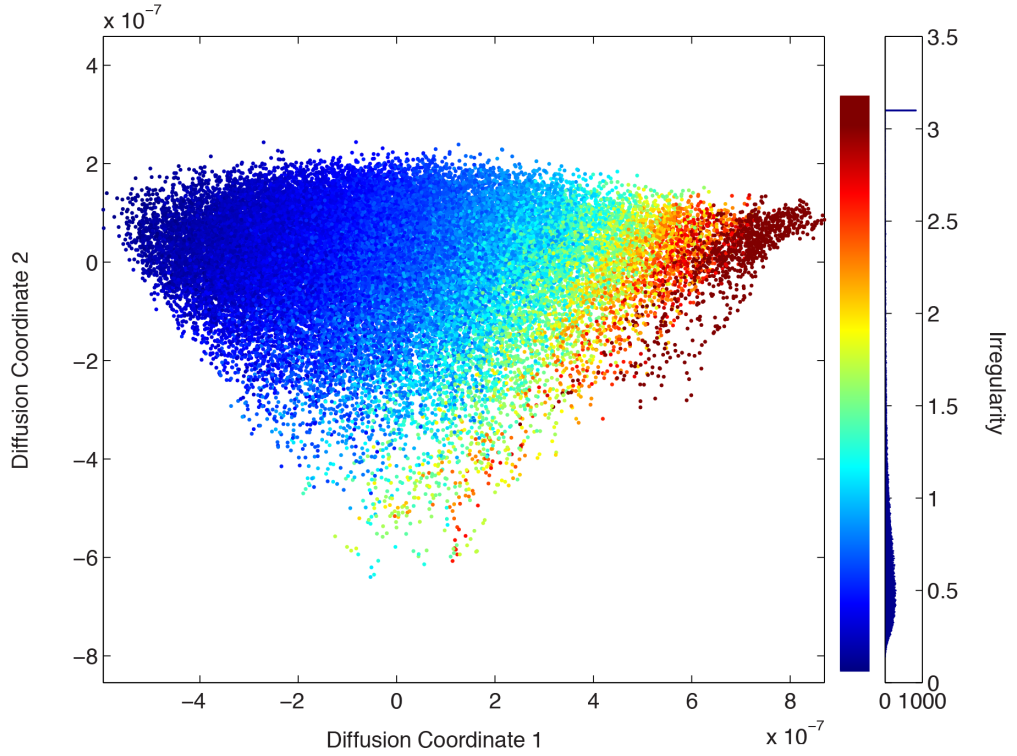


Figure 4.17: **Irregularity distributed over DM embedding.** This figure looks at the distribution of cell shape *irregularity* over the first two Diffusion Coordinates. Each point of the plot represents one cell shape, its position corresponds to its embedding through the Diffusion Maps process, its colour corresponds to its relative *irregularity*. The colour map was generated to display the *irregularity* range, truncated at 3.5 standard deviations either side of the median. Shown in the figure is a histogram showing the *irregularity* distribution over the dataset and a colour bar that corresponds with the *irregularity* value bins. The definition and computation of *irregularity* are explained in section 4.2.1.

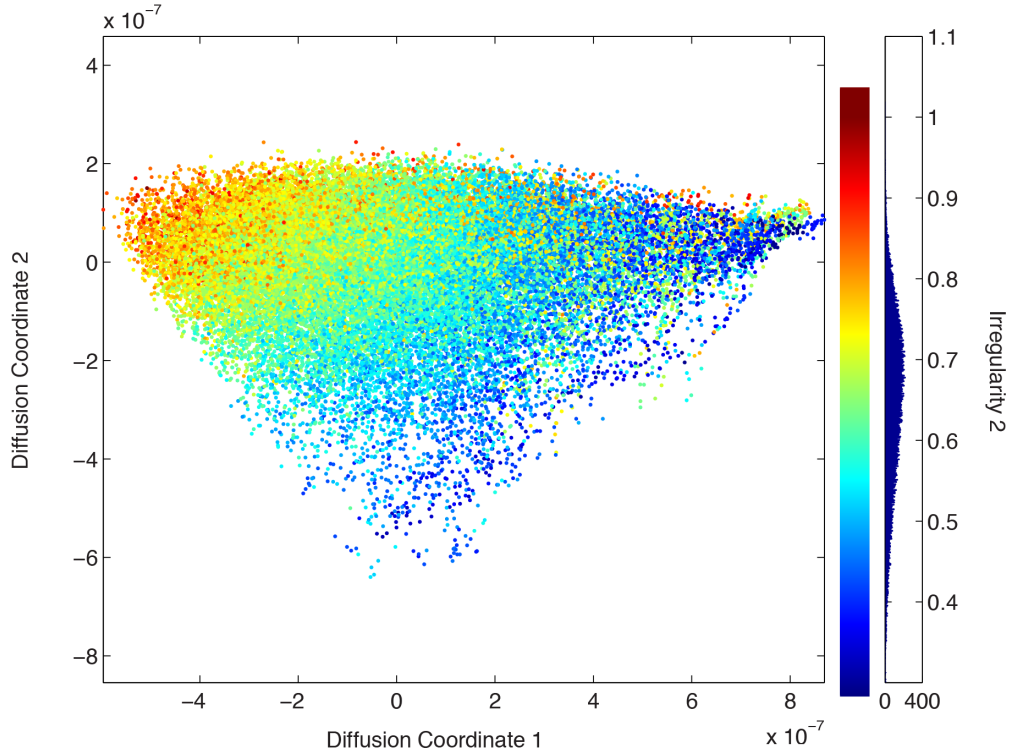


Figure 4.18: **Irregularity2 distributed over DM embedding.** This figure looks at the distribution of cell shape *irregularity2* over the first two Diffusion Coordinates. Each point of the plot represents one cell shape, its position corresponds to its embedding through the Diffusion Maps process, its colour corresponds to its relative *irregularity2*. The colour map was generated to display the *irregularity2* range, truncated at 3.5 standard deviations either side of the median. Shown in the figure is a histogram showing the *irregularity2* distribution over the dataset and a colour bar that corresponds with the *irregularity2* value bins. The definition and computation of *irregularity2* are explained in section 4.2.1.

The figures above clearly reiterate the fact that our newly generated shape representation does indeed represent morphological features present in the data. If a particular feature is well represented by the diffusion coordinates we expect to see good spatial separation according to colour in the figures. For reference, figure 4.7 is a good example of the opposite; the analysis is invariant to the orientation of a cell within the image frame and hence the colours are thoroughly mixed across the embedding.

Our analysis of correlation in section 4.2.2 we highlight the min/max ratio, irregularity and eccentricity as being the features most strongly correlated with the first Diffusion coordinate. This is reinforced by figures 4.6, 4.16 and 4.17, which show very good horizontal separation of colours. Similarly, solidity and symmetry are noted as being highly correlated with the second diffusion coordinate, and figures 4.9 and 4.13 show good vertical separation of colours.

Figure 4.3 (area) looks far less chaotic than 4.7 (orientation), and so perhaps has some more structure. Although it is important to note that the area distribution of is non-uniform. Also, the density of embedded points is not homogeneous. Nonetheless, it appears that cells with smaller area sit in the top left corner.

Figure 4.4 (major axis length) shows very strong spatial separation, and understandably looks similar to 4.14 (max distance from centre). These two distributions, more so than any of the others, appear to separate colours strictly horizontally, that is the feature level sets of these distributions would run almost vertically. This means that the features are strongly represented in the first Diffusion coordinate, which means that the Diffusion Maps algorithm has determined that these features represent a significant amount of the morphological variability present within the dataset. Also the features are independent to the second coordinate.

Figure 4.5 (minor axis length) shows some separation most noticeably in the second Diffusion coordinate. The separation is not perfect and there is some mixing, so while this feature is important it is not the defining characteristic that dictates arrangement.

There is a noticeable small cluster of points around the point  $(0, 1)$  in figure 4.5 (minor axis length), that are red or orange and so have a much higher minor axis length than the others around them. This distinctive cluster is preserved in figures 4.3 (area), 4.4 (major axis length), 4.8 (area of convex hull), 4.11 (perimeter), 4.14 (max. distance from the centroid) and 4.15 (min. distance from centroid). The common thread here is that these features are not normalised by area, and (as is evident from figure 4.3), these cells clearly have much larger area than the other cells with similar shapes. The fact that these points are noticeable indicates that,

in general, cell area is relatively preserved in relation to shape, however these points are an anomaly for which there may be a biological explanation. These points are not visible in the plots of features normalised by area in some form, e.g. figures 4.8 (convex area), 4.9 (solidity), 4.10 (extent), 4.12 (circularity), 4.13 (symmetry), 4.16 (min./max. distance ratio), 4.17 (irregularity) and 4.18 (irregularity2).

One of the most visually striking plots in this section is figure 4.12 (circularity). This figure shows very good separation of colours; there are clear bands across the point cloud dominated by distinct colours. This suggests that circularity is a feature very well reflected with our shape analysis framework, and is also a significant feature in the dataset. Interestingly the level sets of this feature appear to run diagonally, so the feature is somehow important to both Diffusion coordinates.

It is also interesting that while many of the features are similarly distributed across the Diffusion coordinates, they are clearly different distributions (for example, try to guess the angle of the levels in each plot). This emphasises the power that Diffusion Maps has to coordinate the many features that are all jointly responsible for the intrinsic variability within a dataset and to present that information succinctly.

## Chapter 5

# Turn Prediction

### 5.1 Introduction

By simply looking at a still image of a person it is possible to guess whether they are moving or standing still. This is because our familiarity with the mechanisms by which a human can move allows us to predict the intended movement. In this chapter we investigate the extent to which the same can be said of images of migrating epithelial cells. The intention is that, with this work, we can begin to identify the morphological cues corresponding to different mechanisms of migration. The task we set ourselves is to detect turns in each cell's track by looking at the morphological information alone.

In literature, the morphological feature most associated with cell migration is cell polarity, as the cell needs to be able to establish a distinct front and back. In fish keratocytes this is a large fan-like lamella at the front and a thicker cell body at the rear [Abercrombie et al., 1970]. *D. discoideum* cells become elongated and wedge-shaped when responding to high chemotactic stimulation [Tweedy et al., 2013]. Directionally persistent migration in RPE cells has been linked to the maintenance of elongated tails [Theisen et al., 2012].

We start with the following model for RPE cell turning, which assumes a cell maintains directional persistence when it has a prominent front and tail, and hence to achieve a turn the cell must first lose its tail, then choose a new front and new tail to head in a new direction. We describe this morphological behaviour in terms of polarisation; i.e. a cell with a prominent front and tail is called polarised, then the cell depolarises as it loses its tail, then must re-establish a polarity before moving in a new direction.

Thus, if we can detect these depolarisation-repolarisation sequences in the morpho-

logical space, we can predict that there will be a turn in the track at the time of repolarisation. So, if cellular behaviour fits this model consistently we should be able to predict a turn from morphological features alone.

Our BAM based framework, discussed in chapter 4, readily presents a method for distinguishing polarised cells from non-polarised cells; figure 4.1 shows that round cells are most prevalently found at the end of the point cloud with a low 1st Diffusion coordinate, with cells becoming more elongated as this coordinate increases.

A quick look at example RPE1 tracks will suggest that this idea has potential, as, in many cases, the rounded regions match very neatly with a turn in the cell's path. However, it is also clear that RPE cells do not always follow this behaviour mechanism, as they both achieve turns through other mechanisms and repolarise without directional change.

A full model that describes all possible behaviours would constitute a large amount of work, and so is beyond the reach of this thesis. So, in this chapter we limit ourselves to developing a model that describes only this repolarisation-based mechanism. We then develop a framework for estimating the prevalence of this behaviour in our dataset, by checking each cell track against our model.

We continue this section with an outline of the pipeline we have developed for identifying cells that do turn through the repolarisation mechanism. In section 5.2 we introduce the tools with which we analyse the morphological and migrational behaviour, and describe how we train those tools to our dataset. In section 5.3 we present our findings, determining how predominantly this mechanism is used in RPE1 cells, and how accurately we find the location of the turns when this mechanism is used. Finally section 5.4 discusses the successes and limitations of the work in this chapter and suggests the potential for further work.

### 5.1.1 Proposed Pipeline

Our goal, when we examine an individual cell track, is to determine the extent to which the cell's behaviour fits the depolarisation-repolarisation model for directional changes in cell migration. We propose the following two-step algorithm (outlined in figure 5.1).

The first step involves looking at the morphological information for the cell track and, based on this information, making a prediction about where there are turns in the track.

The second step involves looking at the migratory information and examining how well this prediction fits, which means checking the predicted turns to see if they are in fact turns, and checking the stretches between turns to see if these stretches are

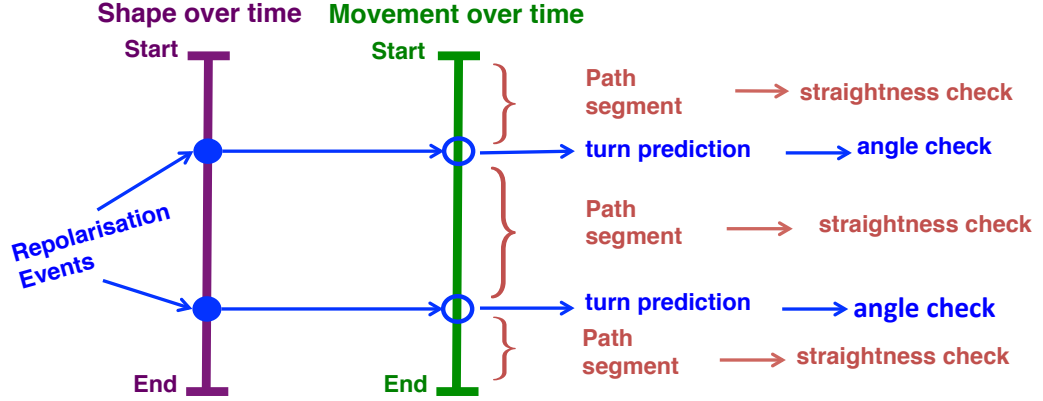


Figure 5.1: **Track analysis pipeline.** This diagram illustrates the algorithm developed in this chapter, for determining how well the repolarisation mediated turning model fits a given cell’s track. The algorithm first analyses the dynamic morphological information of the cell and attempts to detect repolarisation events. From the repolarisation events it then makes a prediction about where the turns in the migratory track are likely to be. We then assess how well this track fits our model by testing whether the turn predictions are correct and also testing whether the path segments between detected turns are straight (i.e. whether or not there are missed turns in this segment).

straight.

In this way, each track contributes four statistics; the number of correct turn predictions, the number of incorrect turn predictions, the number of straight path segments and the number of non-straight path segments. The output of this framework will be the relative sizes of these quantities, which will indicate the prevalence of this mechanism in our dataset.

## 5.2 Development

### 5.2.1 Morphological Analysis

A hidden Markov model (HMM, introduced in section 2.7) is a statistical model for time series data where each observable data point is treated as an emission from a distribution dependent on a hidden state. What is important, though, is that the active state may change after each emission, with a certain transition probability. Often, the task is to view a data sequence and predict the most likely sequence of states that would produce such a data sequence, and hence label or classify each point in the sequence. The benefit in this model is the fact that each data point is classified in context and not in isolation.

The simplest HMM we could devise for use in detecting repolarisations from morphological data within our framework would be a two state model, corresponding to polarised and depolarised. Figure 5.2 (top) shows an example of how one might create such a two-state HMM within our framework. We selected a number of cell tracks that followed the repolarisation mechanism closely and labelled each frame as polarised or depolarised. Then for each labelled set, we fitted a 2D-Gaussian distribution to the Diffusion Maps representation of those shapes (as created in 4.1, only for the first two coordinates); these became the emission distributions of the HMM. From this model we can predict turns to be at the time when a cell transitions from depolarised to polarised.

We also investigated a model with 4 states that correspond to polarised, depolarised and two intermediate states (depolarising and repolarising), see figure 5.2 (bottom). Note that the two intermediate states overlap greatly on the shape representation space, however they have vastly different transition probabilities so they are easily distinguished in context. To make a prediction about the locations of turns we look for sequences in the track that are classified as going from green to blue to red, or depolarised to repolarising to polarised. Then the repolarising (blue) section becomes our predicted turn location<sup>1</sup>. We found that the 4 state model provided a more robust method for predicting turns, and so made use of this in further analysis. Note that the distributions in figure 5.2 do not cover the whole shape space. This reflects our earlier point that many of the cells do not follow the repolarisation model; sometimes through alternative dynamic behaviour, but also through achieving alternative shapes. Given a particular data sequence, it is possible to compute the probability that the data came from this HMM, so we can use this as a tool to detect when cells are not behaving according to our repolarisation based model.

Figure 5.3 gives an overview of one of the tracks used for training. This track was identified as being a good example of the repolarisation mechanism for turning. The figure highlights 8 frames corresponding to important stages of the migratory process, including the repolarisation stage (in blue) which we intend to use as a turn predictor. For these 8 selected stages, the figure displays the image fields, the segmented cell outlines as well as the trajectories of the centre of mass in the image field and the shape representation of the cell track through the Diffusion Maps embedding.

---

<sup>1</sup>In most cases each predicted turn location is one frame. Occasionally, the HMM will label the track as ‘repolarising’ (blue) for more than one contiguous frame. The prediction can simply be interpreted as the cell having changed direction over the course of these frames. We don’t think this devalues the framework over one that strictly places turn predictions at single locations.



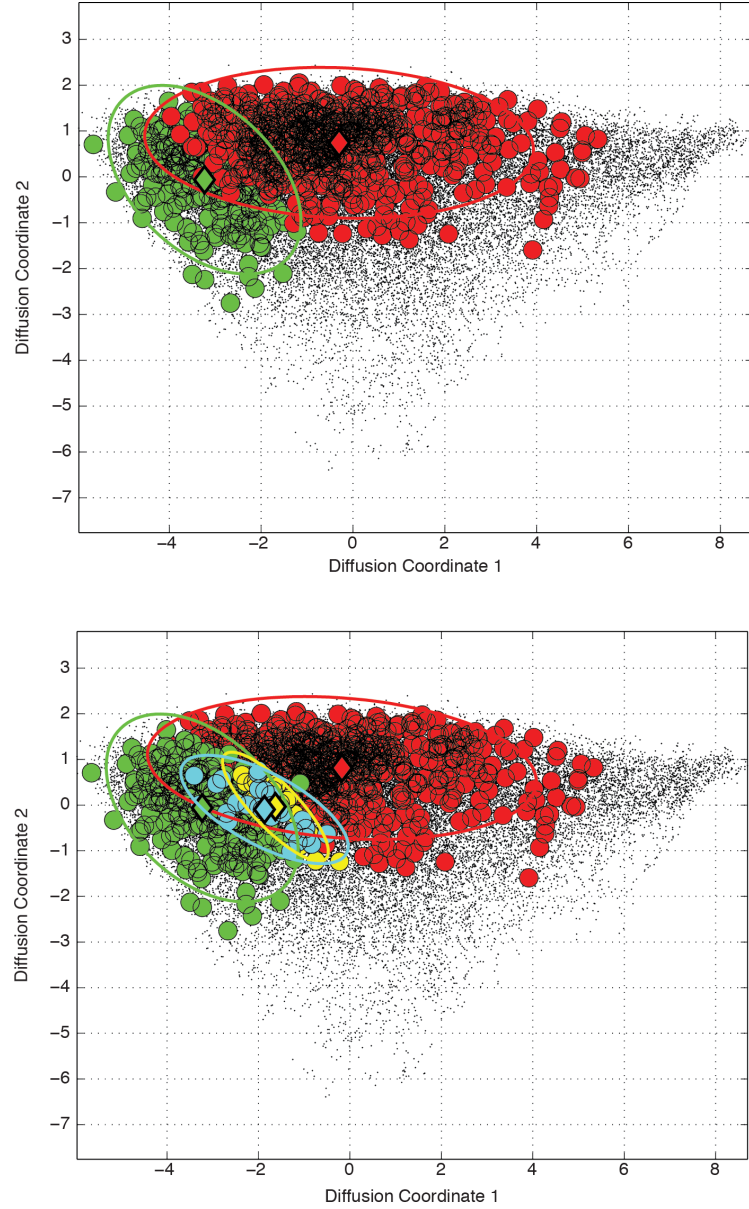


Figure 5.2: **Shape representation for Detecting Repolarisation Events.** These plots show the Diffusion Maps embedding of our set of RPE1 cells, as determined in Chapter 4. Figure 4.1 describes the kinds of shapes represented in different parts of this embedding. Highlighted on each plot are points representing cell outlines in tracks chosen to train a hidden Markov model for cell repolarisation. Each labelled set of points trains a 2D Gaussian, displayed in the plots by a diamond at the mean and an ellipse representing the covariance. *Top:* In the top figure we are training a 2-state model corresponding to polarised (red) and depolarised (green) cells. *Bottom:* In the bottom figure we are training a 4-state model corresponding to polarised (red) and depolarised (green) cells and two transition states; depolarising (yellow) and repolarising (blue).

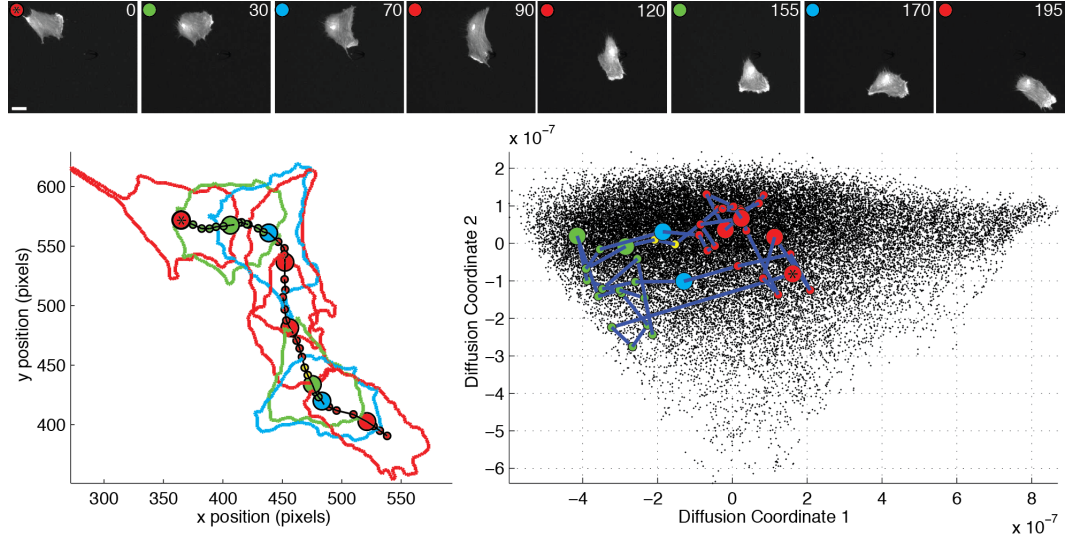


Figure 5.3: **Example cell track labelled for training.** The cell shown in this figure was chosen as a good example of a cell track going through depolarisation-repolarisation cycles, each corresponding with a turn in the cell's path. Eight frames, chosen from the track, are shown, illustrating key stages of the cycle. Time is given in minutes, the scale bar corresponds to  $50 \mu\text{m}$ . All 43 frames in the track were classified into one of four categories representing important stages of the mechanism; red, green, yellow and blue correspond to polarised, depolarised, depolarising and repolarising respectively. These were used to train the hidden Markov model for modelling this mechanism. The bottom left plot shows the path of the 'centre of mass' of the cell through all frames, with enlarged dots and cell outlines on the 8 selected frames. The bottom right figure shows the trajectory of the cell outlines through the Diffusion Maps based shape representation (see section 4.1). The asterisked red dot corresponds to the first frame in the sequence. In our framework, we are predicting that the cell turns will coincide with repolarisation (blue) events.

### 5.2.2 Migrational Analysis

At this stage, for all of our cell tracks, a number of frames may be identified as being predicted turn locations. This prediction has been made using morphological information alone so the next task is to examine the accuracy of the prediction by checking it against the migratory information. The migratory information is represented by the path of the centroid over time (see section 2.8). Two checks must be made to examine the prediction. Firstly we perform an angle check to see if there is a turn where one has been predicted. Secondly we perform a straightness check to ensure there is no turn in segments where none have been predicted.

#### Angle Check

We partition the track into predicted turn points and the path segments between them. To check a given predicted turn location we examine the subsequence of points in the turn location and the path segments immediately before and after. Let the track subsequence be denoted as  $(z_1, z_2, z_3, \dots, z_N)$ , where each  $z_i$  is a planar point  $(x_i, y_i)$ . Let  $a$  and  $b$  be the indices within the track subsequence such that for  $a \leq i \leq b$ ,  $z_i$  is predicted to be a turn location. Note, in most cases  $a = b$ , however it is not required.

We determined that it was necessary to perform two angle checks for each subsequence, one local check and one distant check, defined as follows. The subsequence passes the distant angle check if the angle between the vectors  $(z_a - z_1)$  and  $(z_N - z_b)$  is over  $25^\circ$ . The subsequence passes the local angle check if the angle between the vectors  $(z_a - \bar{z}_a)$  and  $(\bar{z}_b - z_b)$  is over  $40^\circ$ , where  $\bar{z}_a$  is the mean of the set  $\{z_i | \max(1, a - 10) \leq i \leq \max(1, a - 6)\}$  and  $\bar{z}_b$  is the mean of the set  $\{z_i | \min(N, b + 6) \leq i \leq \min(N, b + 10)\}$ .

Figure 5.4 illustrates why both are necessary, since for some tracks a path segment will curve significantly, causing the distant check to incorrectly fail, and for other tracks a delay occurs after repolarisation, causing the local check to incorrectly fail. So, a turn prediction is marked to be correct if it passes either of these checks.

The thresholds were trained using a labelled set of training sequences, each containing one repolarisation-based turn. Track sequences that were identified as curving after or before the turn were excluded from the training of the distant check, track sequences that were identified as delaying after the turn were excluded from the training of the local check.

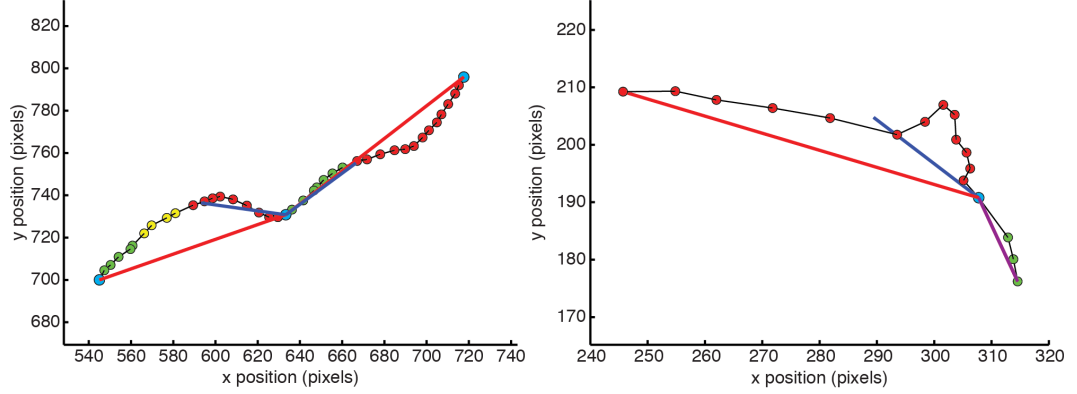


Figure 5.4: **Angle check difficulties.** This diagram illustrates the necessity for a two-stage angle check. The figures show the path of the centroid of a cell over time. The centroid is coloured according to its morphology based HMM classification (see section 5.2.1). Red and blue lines represent the lines used for angle checks (blue for local, red for distant, see section 5.2.2), the magenta line represents coincident red and blue lines. In the *left* figure the cell turns at the point of a repolarisation, but then curves back out to its original direction. In the *right* figure there is a delay between the repolarisation and the cell moving in it's new direction.

### Straightness Check

We have already partitioned the track into predicted turn points and the path segments in between them. To check for the absence of turns in each path segment we perform a straightness check as follows. We first find the linear fit of the path segment, then compute the perpendicular distance of each point to the fitted line. Denote these perpendicular distances by  $(d_1, d_2, d_3, \dots, d_N)$ , where  $N$  is the number of points in the segment. Then we assess the straightness by computing

$$D = \frac{1}{L} \sqrt{\sum_{i=1}^N d_i^2}, \quad (5.1)$$

where  $L$  is the length of perpendicular projection of the whole path segment onto the fitted line. We then say the segment is straight if  $D$  is less than 0.175. This threshold was optimised to discriminate a labelled training set of straight and non-straight path segments (all identified as not following a repolarisation mechanism).

### 5.2.3 Turn Prediction Accuracy

The framework as outlined above is designed to determine how often our RPE cells follow the repolarisation-based mechanism for turning. Moreover, for the tracks

identified as following this mechanism, the algorithm produces a prediction for the location of the turn. We now present our method for assessing the accuracy of the turn predictions.

To do this, we restrict ourselves to subsequences of cell tracks that contain a turn prediction which passes the angle check surrounded by path segments which pass the straightness check. We define the true location of the turn to be the point that has the highest perpendicular distance to the straight line that connects the end points of the subsequence<sup>2</sup>. Then we measure the time delay between the predicted turn location and the true turn location. For contrast we also measured the time delay between the predicted turn location and two other landmarks; the midpoint of the path segment before the true turn location, and the midpoint of the segment after. We report the results in section 5.3.

### 5.3 Results

Applying our final algorithm for turn prediction (the development of which is described throughout Section 5.2) to a set of 440 cell tracks (disjoint from any tracks used in training), we get the following results:

- The algorithm detected 460 repolarisation events, of which 339 passed the angle check and 121 failed.
- Partitioning the tracks by repolarisation events yielded 889 path segments, of which 500 passes the straightness check and 389 failed.

The first result above suggests that 78% of the repolarisation events corresponded with an angle change. To look more thoroughly at the distribution, figure 5.5(A) shows the full range of angles made at the detected repolarisation events (defined as distant angles as in section 5.2.2), and for comparison figure 5.5(B) shows the distribution of angles made at mid-points of path segments between detected repolarisations events. Although repolarisations do not always correspond with a significant directional change, we can see that the range of angles is much broader for directional changes at a repolarisation event than in the middle of other path segments. So, a repolarisation permits a cell to turn, but does not cause a turn, this is why the angle variance of a depolarised cell that repolarises is higher than one that stays polarised.

---

<sup>2</sup>This ought to be a robust method for finding the turn location, because we are only looking at sections comprised of two straight segments.

If we assume that each non-straight path segment represents a turn through a mechanism that is not repolarisation, then we can say that approximately 47% of turns are through repolarisation. This is a somewhat naive interpretation of the results, but without a thorough understanding of all possible mechanisms of migration, and how to identify each of them, this sort of statistic is difficult to generate.

Figure 5.6 shows 6 example tracks that contain a repolarisation event that passes the angle check, surrounded by paths that pass the straightness check. In each case the blue point (the predicted turn location) is on or near the apex the of corner. These are examples of cells seeming to follow the repolarisation model for turning. Figure 5.7 shows 5 tracks chosen for discussion; in each case one of the checks has failed. We emphasise that these ‘failures’ do not represent flaws in the algorithm but rather they represent identification of cell tracks following biological behaviour different from our repolarisation model. Tracks A, D and E contain paths that fail the straightness check. Tracks B and C contain turn predictions that fail the angle check. An interpretation of the mechanisms displayed in these tracks will be given in section 5.4.

To analyse turn prediction accuracy we investigated 155 ‘straight-corner-straight’ path subsequences from 131 different cells. We computed that in more than 16% of the cases the repolarisation event was the closest time point to the location of the turn, in more than 37% of the cases the repolarisation event was in the top 3 closest points and in more than 61% of the cases the repolarisation was in the top 5 closest points. Figure 5.8 shows (in the centre plot) the distribution of the

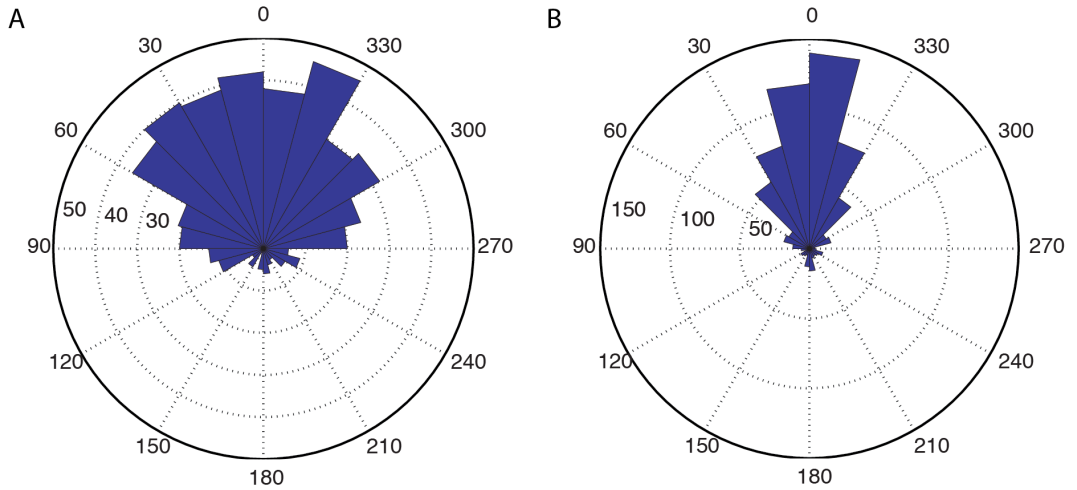


Figure 5.5: **Angle distributions.** The distribution of directional changes measured at (A) detected repolarisation events and (B) path segment midpoints.

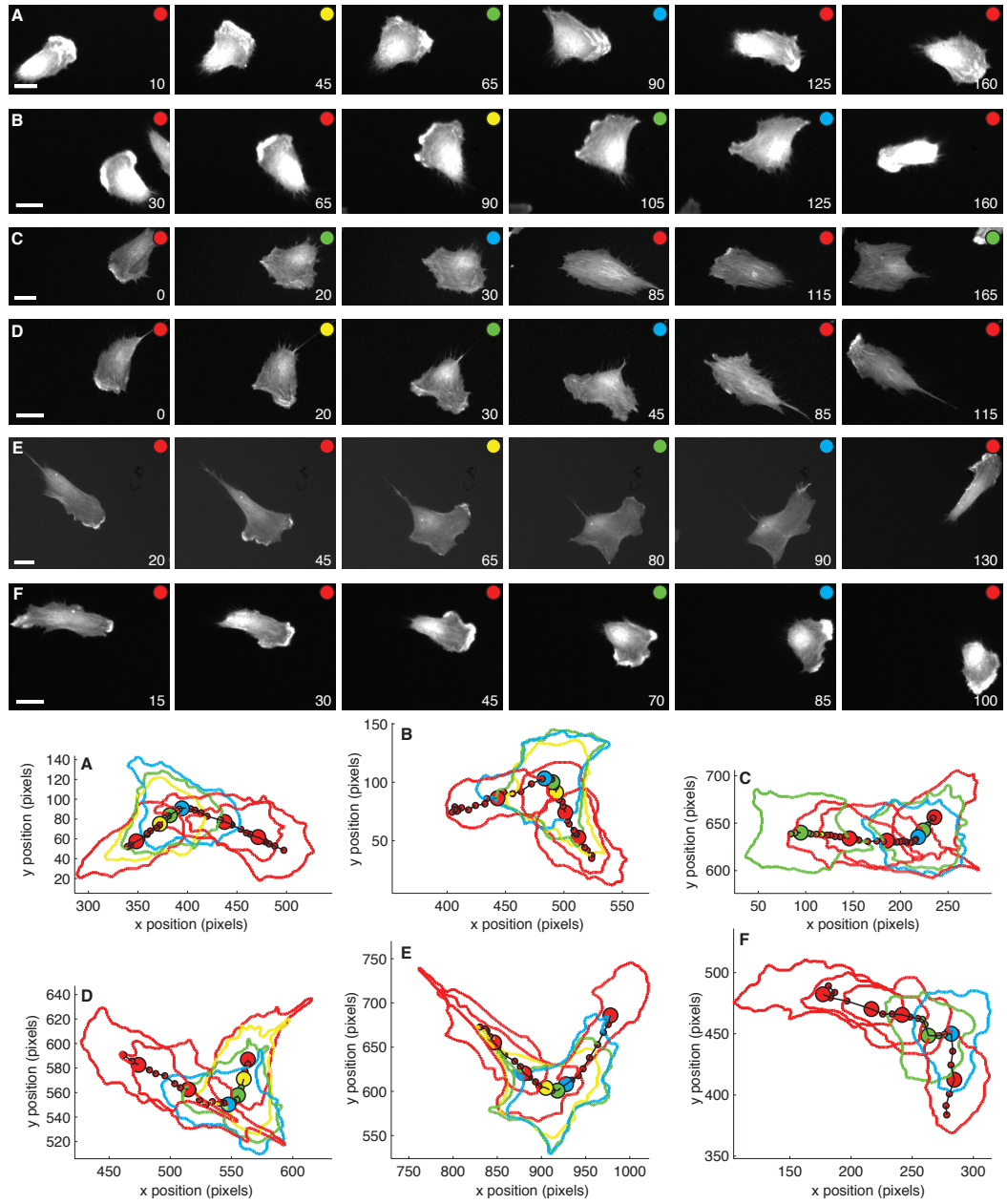


Figure 5.6: **Example repolarising tracks.** Six cell tracks (A-F) each demonstrating a turn using the repolarisation mechanism, and each correctly identified with our HMM track analysis. For each track, six selected frames are shown, with a scale bar of  $50\mu\text{m}$ , times shown in minutes and HMM shape classification shown by a coloured circle. For each track, we also show the cell outlines superimposed with the path of the 'centre of mass'.



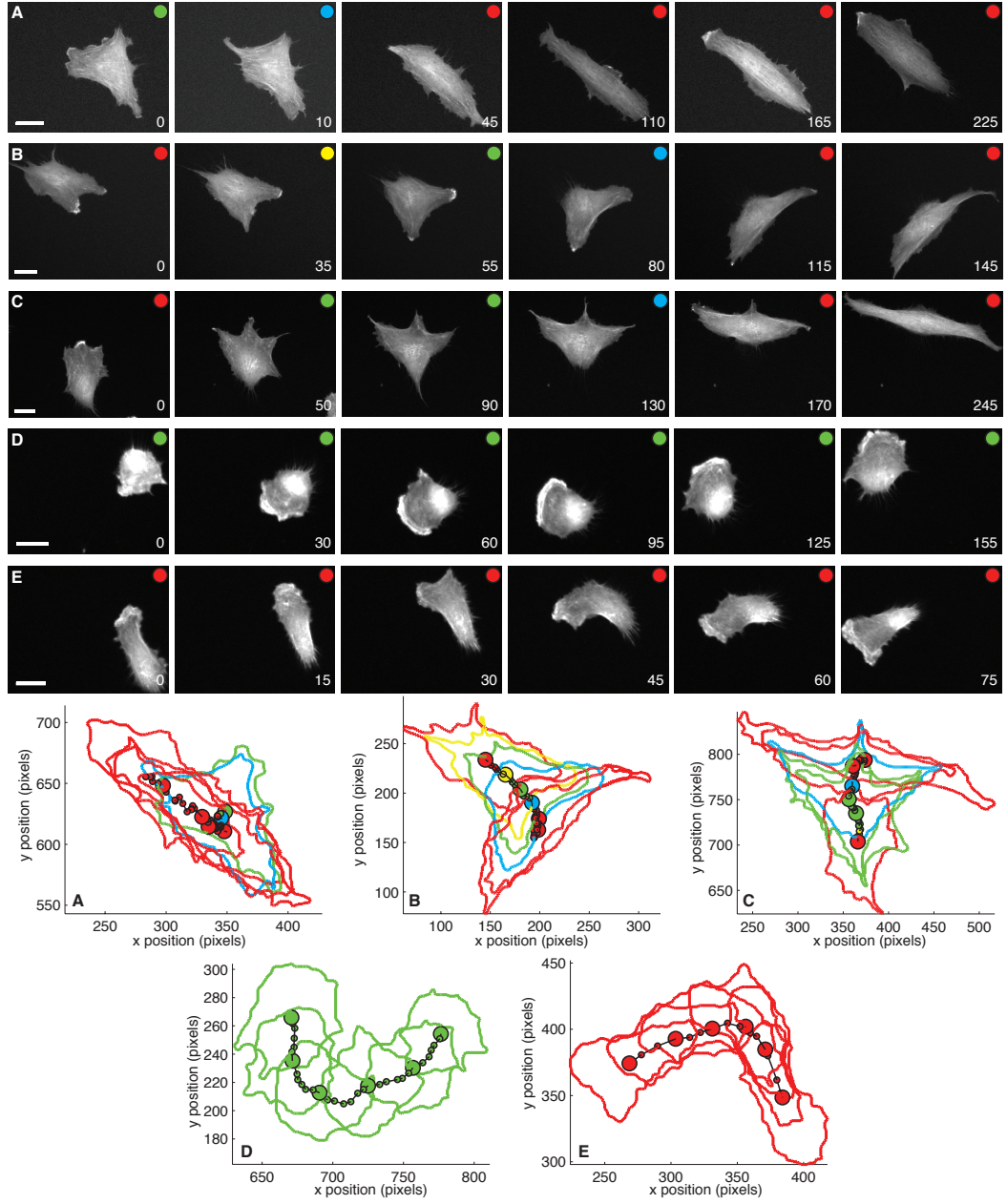


Figure 5.7: **Example alternative tracks.** Five cell tracks (A-E) each demonstrating a mechanism unexplained by our simple repolarisation model. For each track, six selected frames are shown, with a scale bar of  $50\mu\text{m}$ , times shown in minutes and HMM shape classification shown by a coloured circle. For each track, we also show the cell outlines superimposed with the path of the ‘centre of mass’.



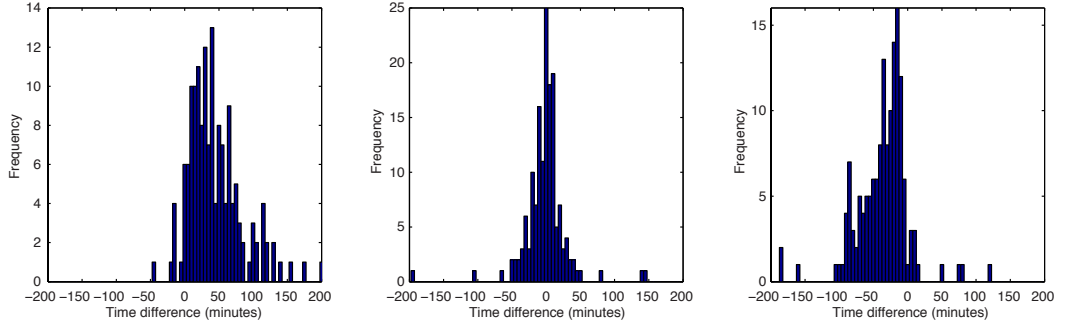


Figure 5.8: **Turn prediction accuracy.** This figure examines the precision of the morphology based turn prediction in tracks identified as following the repolarisation mechanism. 155 tracks are examined. For each track the true location of the turn is identified from the migratory information. Each histogram shows the distribution of time differences between the predicted location of turn and (*left:*) the midpoint of the track before the true turn, (*centre:*) the true turn or (*right:*) the midpoint of the track after the true turn.

time disparity between the location of the turn and predicted turn. For comparison we also show the distribution of the time disparity between the predicted turn and the points closest to the midpoints of the linear fits of the surrounding segments.

## 5.4 Discussion

This work supports the idea that RPE1 cells can achieve a turn through a repolarisation. It has become clear that this is not the only mechanism that these cells will use to turn, but it is the mechanism that singly explains most turning events. We have also shown that it is possible to quantitatively model the morphological behaviour displayed during a repolarisation. We accept that with more time and work a more complete model could certainly be created, however we believe that this preliminary work shows that it is possible and confirms that morphological data does have predictive power over migrational data. This is important since it means that there is a measurable relationship between shape and movement, as we originally sought to show. Further study of the intracellular mechanisms involved in cell dynamics could be aided by a model of this relationship.

We believe that this chapter gives a successful example of using our Diffusion Maps representation of cell shape to perform a task. This work is our first attempt at modelling *dynamic* cell shape data with our Diffusion Maps representation. We believe it highlights the potential use for this representation in complex applications.

Figure 5.7 shows some example tracks that do not strictly follow our repolarisation model. The behaviour seen in tracks A-C of figure 5.7 can be described as a bifurcation. A bifurcation is a front-induced process where the leading edge splits, protruding in both sideways directions, one of which later becomes the tail and the other the front. This does not necessarily conflict with the repolarisation mechanism for turning, since it still involves a break-down and re-establishment of polarity, but it can contribute to behaviour that is unexpected in our model. Most often, the issue is a delay after depolarisation before the cell moves in its new direction.

In track A the cell goes through the bifurcation process; after repolarising, the cell spends approximately 100 minutes deciding on the new direction, before moving away from the location of the turn. This means that the path after the repolarisation event is not straight. In track B the cell drifts while repolarising, so much so that the cell soon touches the image field edge; the cell’s track is cut short and we do not see which direction the cell eventually moves in. The result is that the cell’s centroid moves in the original direction and the track is identified as not turning during the repolarisation. Track C shows an example of when a cell grows in its new directions before retracting its tail; this results in bizarre cell shapes that lie far from the emission distributions of our HMM. Interestingly the algorithm still manages to correctly identify the repolarisation, even though the shapes are very different from the training set. It would be interesting to investigate this behaviour more closely and try to develop a way to predict (through morphology) whether, after repolarisation, a cell will bifurcate or move off in another direction.

Tracks D and E of figure 5.7 show other examples of cells not following our simple mechanism. The cell in track D manages to turn a corner while staying depolarised, whereas the cell in track E turns a corner remaining polarised. There are certainly other mechanisms in play beyond repolarisations and bifurcations; we suggest that the framework presented in this chapter can help in the identification of other mechanisms. Automatic detection of cell tracks whose migratory behaviour does not fit the predicted behaviour based its morphological information will help identify new mechanisms to investigate.

One perceptual difficulty with the problem statement of ‘turn prediction’ is that there is no formal definition of a ‘turn’. It would be nice to be able to define a turn solely based on the path of a mid-point, however, cells will frequently protrude and deform laterally whilst maintaining a persistent direction, this makes it very difficult to rigorously distinguish a ‘wiggle’ in a cell’s path from a true turn that may be shallow. For a human observer to distinguish between these two, he/she would need to view the original cell image sequence to perceive whether the turn

was a deliberate, mechanically-induced turn or whether it was indeed a sideways fluctuation. Therefore it might be useful to actually define a turn as a sequence of morphological stages. Working towards a morphological definition of turning would encourage an understanding of the mechanism by which a cell achieves a turn rather than simply understanding the outward behaviour.

With a full model, an immediate application could be to analyse data from perturbation experiments and automatically identify segments of cell tracks that correspond to known mechanisms. It would then be possible to quickly look for differences in the migratory mechanisms on a population wide level. Theisen et al. [Theisen et al., 2012] identify Kif1C as being important in cell tail stabilisation and report that silencing this protein reduces the lifetime of each cell tail, resulting in rapid depolarisations. With our framework we ought to be able to detect these rapid depolarisations and quickly measure their frequency on a population-wide level. We would also be able to examine the morphological data of a cell as it turns and look to see if there are any other mechanistic differences between the siKif1C and the wild type cells. If one had a suitably automatic segmentation algorithm, one could run high-throughput experiments to look to see if there are genes that are particularly responsible/important for any one of the identified migratory mechanisms.

Other applications for this body of work include cell segmentation. One major challenge for automatic segmentation for image sequences of migrating cells is occlusion, as cells will occasionally bump into or climb over one another. With this framework, one could examine the cell outlines as they are before a suspected occlusion, and make a prediction about the shapes and locations of the cells during an occlusion. With translucent cells, where the individual outlines are visible during an occlusion (but difficult for an algorithm to find), this framework could give an initial estimate for the segmentation algorithm to work with.

Another application for this work lies in generating synthetic data. If one had synthetic path data, it would be possible to use a framework similar to the one presented here to generate shape data to accompany the path data, and so create full-image scenes. Synthetic data generation is valuable within the image computing community as it represents unlimited data where the exact ground truth is known.

## Chapter 6

# Discussion and Conclusions

### 6.1 General Discussion

This project began in a very exploratory fashion. The suggestion was to apply a machine learning technique, which had seen success in quantifying shape [Rajpoot and Arif, 2008], to biological images and see what could be learned. RPE cells presented an interesting case to examine as their widely varying morphologies had been difficult to model in the past, and clearly their shape dynamics were integral for their function and migratory behaviour. So we began with the hypothesis that there would be a quantitatively measurable dependency between morphology and migration in these cells in a way that would not have been possible with only previous simple techniques in quantifying shape information.

Preliminary work had shown the potential of the shape space learning framework, however we needed to improve the algorithm. The crucial component being the shape similarity measure as this is the piece that most closely interacts with the dataset. Also it became apparent in many of the preliminary experiments that our datasets were not large enough. This was clear because we could see ‘bizarre’ shapes that would appear in only one cell (and remember that this cell will span a number of frames) per dataset, and these were treated by our learning algorithm as entirely different clusters, far removed from the other data points. The dataset did not include the full spectrum of shapes leading to these bizarre shapes because they are substantially rarer than the simple shapes, however we knew from discussion with biologists familiar with the data that these weird shapes are legitimate actions of the cell. So during the project it was necessary to gather a much larger dataset, and develop an algorithm that could handle it.

## 6.2 Shape Analysis

### 6.2.1 Shape Space Learning with Diffusion Maps

The major contribution of this project is the presentation of a framework for generating a quantitative representation of shape. The data analysed in this work is a set of cell outlines of migrating human RPE cells, however the framework can be used for any set of simple closed planar curves. The framework is designed to learn the shape properties that are most responsible for the observed variation in the dataset, and so needs no prompt or supervised labelling, but can be applied to a dataset where the intrinsic degrees of freedom are not yet fully known or understood. The framework has been successful in representing a contiguous dataset where the observed variation occurs as a continuous spectrum and not as discrete clusters. The framework allows for efficient implementation to analyse large datasets, which may be necessary in cases where subpopulations are a significant minority.

The output of our shape representation framework has mainly been presented as simply the top two Diffusion coordinates. In many cases this is simply to make visualisation easier, but later it seemed as though the top two coordinates were sufficient to model the repolarisation-depolarisation behaviour that we were examining (chapter 5). It is possible that more information is held in the other coordinates that may aid in other applications including the modelling of other migratory mechanisms.

One of the biggest difficulties when developing this framework was validating its performance. This was because there was no ground truth against which to compare our shape representation, ~~since shape is not intrinsically quantitative~~. For this reason we invested a lot of effort into developing ways to visualise the features that our representation framework captures (see Chapter 4). We believe this is appropriate since, in the absence of a deeper truth, human perception provides the most authoritative assessment of shape. However, while visual inspection can confirm success in a binary sense, it does not allow comparative performance assessment for competing algorithms. This means tasks such as parameter optimisation become very difficult.

Given that shape representation is unlikely to be performed in isolation, but rather to be incorporated into a model for some downstream analysis, we recommend that optimisation can be achieved based on performance at a later stage of the analysis. However, since our downstream analysis (of migratory behaviour) was exploratory in nature, and again we had little understanding of the underlying truth, we struggled with this optimisation.

One parameter that fell victim to this difficulty was the similarity kernel bandwidth (see equation 3.1 or section 3.3). This parameter is responsible for interpreting the contextual idea of near versus far, it is important in many machine learning contexts and is still an open problem in the community. In our context this parameter can have a dramatic affect on the geometry of the generated shape representation. Section 3.3 describes the few methods that we used to try to generate this parameter, but ultimately we found that using the median of measured distances gave a satisfactory result for our further purposes. We suggest that finding a way to robustly select an appropriate value for this parameter would be a very worthwhile extension to this work.

We suggest two more extensions to our shape analysis framework that follow other efforts in extending the Diffusion Maps framework in literature [Coifman and Lafon, 2006b; Rabin and Coifman, 2012]. That is, the inclusion of out-of-sample data to a learned distribution and the generation of synthetic data from a learned distribution. We believe that both of these processes could be readily integrated into the framework and would yield very worthwhile extensions.

We began work on the out-of-sample extension, but due to time constraints we omitted this work from the thesis. The shape representation framework we presented here produces a low dimensional representation of a given dataset of shapes; the out of sample extension takes data from outside of the original dataset and embeds it into the generated representation space. This can be achieved using a combination of Geometric Harmonics and Laplacian Pyramids as described in [Coifman and Lafon, 2006b] and [Rabin and Coifman, 2012]. This is could be useful when examining data from perturbation experiments or other distinct populations in the context of the learned wild type distribution. We performed this out-of-sample embedding with data collected from Kif1C interfered RPE cells. Kif1C interference is known to inhibit the ability for RPE cells to maintain cell tails [Theisen et al., 2012], and correspondingly our shape analysis placed these cells predominantly amongst the shorter cells (low first Diffusion coordinate, see figure 4.1). We omit these results from this thesis, since this analysis is incomplete, however this would be interesting to pursue, especially once more is understood about the dynamic behaviour of unperturbed cells.

The extension can also be used for back projection, i.e from an arbitrary point in the low dimensional coordinate system (within the range of the embedded training points) it is possible to generate a synthetic shape that reflects the relative position of the low-d point. This has many applications, in section 5.4 we discuss the idea of generating synthetic image sequences of migrating cells and the subsequent

benefits. Another benefit would be that we can create shape data for which there is a quantitative ground truth already known, which can be used to train, test and optimise other algorithms.

### 6.2.2 Best Alignment Metric

In this thesis we present development of a novel shape distance measure specifically designed for the comparison of independent simple closed planar curves. This will find the pairwise distance between corresponding pairs of points on the curves after the curves have been mutually aligned, reparameterised and interpolated so as to best emphasise the similarities between the curves. This means it satisfies the specific requirements we have for a shape metric, as outlined in section 3.2.1.

The formulation employs the Fourier transforms of the curves. This allows for explicit computation of the optimum solution to the mutual angle of the curves and circular convolution which accelerates finding the overall optimum solution, resulting in fast computation of this metric.

Understanding the requirements for this metric took a surprisingly large amount of time, and only became clear after significant effort in the wrong directions. Most shape similarity metrics in literature rely on some internal coordinate system to find mutual alignment. Sometimes this internal coordinate system relates to landmarks in the shape curve, e.g. when comparing the shapes of a person, one should align heads with heads and feet with feet (see [Cootes et al., 1995] for a classic example of landmark alignment with resistors). Other times, the alignment comes from an extrinsic coordinate system, for example in the analysis of local membrane deformation one can look at sequential images of a cell and the frames will be inherently aligned. In our analysis we look for preserved morphological phenotypes across independent cells that have no reliable landmarks, hence we were required to mutually align pairs of cells.

In literature, much of the work allows mutual alignment for shape comparison is based on the elastic distance [Younes, 1998]. This achieves mutual alignment through the consideration of equivalence classes under group actions on the curves. While this theory is elegant and satisfies our requirements, in our experiments we found implementations of this framework (using [Joshi et al., 2007]) to be prohibitively slow. This led to the development of BAM which similarly considers equivalence classes, but can be computed rapidly due to the simple nature of the underlying metric, the  $L^2$ -norm. While the metric may not be as sophisticated as the elastic distance, we believe the output is sufficient for our application, since we are looking for emergent properties of a large dataset.

One potential extension for the use of BAM could be to prealign the shapes before using a more sophisticated metric to help alleviate the computational burden that these approaches require.

## **6.3 Mechanisms of migration**

### **6.3.1 Turn Prediction**

One of the major goals for this body of work was to show that a quantitative interpretation of shape can be useful in seeking an understanding of the mechanisms of migration. In chapter 5 we describe an algorithm that makes predictions about the occurrence of track turns looking only at the dynamic shape information of a cell. We restrict our investigation to a particular turning mechanism, but show reasonable success. This shows that cell shape does have a quantitatively measurable relationship with migration, and that our framework has the potential to investigate it.

Our work in the investigation of cell migration is limited to the repolarisation mechanism for turning, purely due to time constraints. We believe that continued effort in this direction could yield many interesting findings. Firstly there are other identified turning mechanisms that one could model with the same strategy as used in chapter 5. Secondly with an established quantitative model of cell dynamics at this level, it would be very interesting to investigate the effects that genetic perturbations had on the system at the cell behaviour level. A significant benefit of our framework is that our shape analysis is quantitative, this means we can measure statistical significance of any observations and even run high-throughput experiments. These were not previously possible with subjective interpretations of cell shape.

### **6.3.2 Our Migration Analysis in Context**

It is widely agreed in the literature that for a cell to migrate with directional persistence it must establish and maintain polarity (a distinct front and rear) [Ridley et al., 2003]. Most examples in literature attempt to model the mechanisms through which a cell achieves this regulation of cell polarity during periods of persistent migration. Our examination of track turns can perhaps be regarded as the contrapositive; in order for a cell to turn (to cease directional persistence) it must first disengage whichever mechanism is maintaining polarity.

One idea is that protrusion elsewhere than at the front of the cell is suppressed



by an inhibitor, which in only overcome at the front of the cell. Such an inhibitor could be a soluble chemical substance or a mechanical signal. Recently, tension has been suggested to function as a global inhibitor that can only be overcome at the front. Houk et al. apply tension artificially and release tension using ablation experiments to advocate the role of membrane tension as the long-range inhibitor of lateral protrusion in migrating neutrophils [Houk et al., 2012]. Mogilner and Zhu conclude from the work of Houk et al. that tension is both necessary and sufficient to polarise the cell [Mogilner and Zhu, 2012]. An alternative view suggests that mechanical force from contacts to other cells or the substrate promotes polarization and protrusion in the opposite direction to the initial force. Therefore, cells generating drag forces at their tail move with higher directional persistence than cells with perturbed tail dynamics [Theisen et al., 2012]. Likewise, mesendodermal cells that migrate *in vivo* as a collective tissue integrate the forces from neighbouring cells through cell-cell contacts [Weber et al., 2012]. When these cells are dissociated from one another and plated onto fibronectin *in vitro* they lose their unidirectionality and become multipolar. Polarisation and directional migration can be induced by applying forces through C-cadherin-coated beads. Cell protrusion occurs in the opposite direction to the tension applied, supporting the idea that tension at the rear stimulates front-protrusion [Weber et al., 2012].

Our observation that many cells depolarise ahead of a turn fits well with these models, since it is clear that a rear retraction will reduce the tension throughout the cell and consequently end the inhibition of lateral protrusion and/or the promotion of front or back protrusion. Further evidence for the importance of the tail and the rear of the cell being a driving component in the mechanisms of symmetry breaking and migration comes from the careful analysis of the timing of events at the front and rear of the cell during the initiation of cell migration or during repolarisation [Rid et al., 2005; Yam et al., 2007; Cramer, 2010]. This again fits with our observations, since we often see the retraction of the tail before any other morphological change.

Much of the literature on cellular motility focuses on chemotaxis which is often explained purely through front-lead mechanisms [Weiner, 2002; Andrew and Insall, 2007]. Two alternative models explain the directional change of cells subjected to a change of a chemotactic gradient. The compass model suggests that a change in the gradient induces a new protrusion in the direction of the gradient from the site where chemotactic receptors are stimulated the strongest. The informed choice model proposed by Andrew and Insall suggests that the cell makes a number of protrusions and the one that happens to move up the gradient more, is reinforced

[Andrew and Insall, 2007]. We also observe alternative modes of cell turning that do not involve substantial depolarisation. These can either take the form of a front diverting gradually and dragging the cell body behind, the front splitting into two or more protrusions with one becoming the new front and the other the new tail. Both these mechanisms appear to be front-led. Thus the freely migrating cells we observe use a number of mechanisms for cell turning and we expect that the frequency with which each of these mechanisms in its repertoire are used will change depending on the environment. The dynamic shape analysis routines developed in this work will enable the objective analysis of turning mechanisms in perturbation experiments to unravel the molecular mechanisms involved.

## 6.4 Plan for Publication

Our plan for publication of this work is to develop a user-friendly, open-source toolbox for quantitative cell shape representation that allows a user to perform much of the analysis presented in this thesis. The software will take most microscope image formats and take the user through the processes of segmentation, low-dimensional representation, clustering and visualisation. These dataset representations will all be exportable for further quantitative analysis.

We aim to present a biological methods paper that describes our algorithms and introduces our toolbox with examples from our RPE set. We aim to reinforce this work with successful embeddings of other cell lines. We will also discuss our investigation into turning prediction for RPE cells as an example application to dynamic behaviour modelling. The shape representation toolbox will have some basic dynamic behaviour tools, however the toolbox ought to be applicable to many cell lines and varied experiments, so we would encourage dynamic models to be designed to each situation.

# Appendix A

## Best Alignment Metric

### A.1 Best Alignment Metric Formulation

We want to define a distance measure between pairs of shapes (closed curves) that accommodates alignment and cyclic reparameterisation variation, this is not simply a matter of invariance to these things, the metric must find an “appropriate” pairwise alignment. As “appropriate” we are using the pairwise alignment that minimises  $L^2$  (or  $\ell^2$ ) distance. Translation invariance is included by translating each curve so that the mean of the points lies on the origin, appendix section A.2 shows that this translation choice minimises the relevant measurement. This chapter will deal with a finite element approximation to curves in the plane, one can (I’m sure) restate the following work with continuous curves.

Approximate a closed curve by a sequence of  $N$  complex numbers whose mean lies at zero. We then define the shape metric between  $u = (u_1, u_2, \dots, u_N)$  and  $v = (v_1, v_2, \dots, v_N)$  to be:

$$d(u, v) := \min_{r, \phi} \sqrt{\sum_j |u_{j+r} e^{i\phi} - v_j|^2}. \quad (\text{A.1})$$

Indices are taken mod  $N$ . Here,  $\phi$  represents rotation of the plane, and  $r$  is a cyclic shift of the indices (the finite element equivalent of cyclic reparameterisation). We now wish to simplify expression (A.1) to allow for rapid computation.

**Lemma 1.**

$$|a - b|^2 = |a|^2 + |b|^2 - (\bar{a}b + a\bar{b}). \quad (\text{A.2})$$

*Proof of Lemma 1.* First we look at

$$\begin{aligned}\bar{a}b &= (Re(a) - iIm(a))(Re(b) + iIm(b)) \\ &= Re(a)Re(b) + Im(a)Im(b) - i(Im(a)Re(b) - Im(b)Re(a)),\end{aligned}$$

and equivalently

$$\begin{aligned}a\bar{b} &= (Re(a) + iIm(a))(Re(b) - iIm(b)) \\ &= Re(a)Re(b) + Im(a)Im(b) - i(Im(b)Re(a) - Im(a)Re(b)).\end{aligned}$$

This gives us

$$\bar{a}b + a\bar{b} = 2(Re(a)Re(b) + Im(a)Im(b)),$$

so we can now compute

$$\begin{aligned}|a - b|^2 &= (Re(a) - Re(b))^2 + (Im(a) - Im(b))^2 \\ &= Re(a)^2 - 2Re(a)Re(b) + Re(b)^2 + Im(a)^2 - 2Im(a)Im(b) + Im(b)^2 \\ &= Re(a)^2 + Im(a)^2 + Re(b)^2 + Im(b)^2 - 2(Re(a)Re(b) + Im(a)Im(b)) \\ &= |a|^2 + |b|^2 - (\bar{a}b + a\bar{b}).\end{aligned}$$

□

Now, using Lemma 1 we can reduce our metric to

$$d(u, v) = \sqrt{\sum_i |u_i|^2 + \sum_i |v_i|^2 - \max_{r, \phi} \left( \sum_j \bar{u}_{j+r} v_j e^{-i\phi} + \sum_k u_{k+r} \bar{v}_k e^{i\phi} \right)}. \quad (\text{A.3})$$

### A.1.1 Dealing with $\phi$

To simplify the expression  $(\sum_j \bar{u}_{j+r} v_j e^{-i\phi} + \sum_k u_{k+r} \bar{v}_k e^{i\phi})$ , we first note that the terms  $e^{-i\phi} \sum_j \bar{u}_{j+r} v_j$  and  $e^{i\phi} \sum_k u_{k+r} \bar{v}_k$  are complex conjugates. If we fix  $r$  and vary  $\phi$  we rotate each of the terms in opposition and their sum will always lie on the real line (in fact this is a necessary condition for us to take a maximum as we do in (A.3)). Clearly the maximum of the sum occurs when both terms are real and positive (and hence equal), so for fixed  $r$

$$\max_{\phi} \left( \sum_j \bar{u}_{j+r} v_j e^{-i\phi} + \sum_k u_{k+r} \bar{v}_k e^{i\phi} \right) = 2 \left| \sum_j \bar{u}_{j+r} v_j \right|.$$

So now we can reduce the distance metric to

$$d(u, v) = \sqrt{\sum_i |u_i|^2 + \sum_i |v_i|^2 - 2 \max_r \left| \sum_j \overline{u_{j+r}} v_j \right|}. \quad (\text{A.4})$$

### A.1.2 Dealing with $r$

Next we deal with rapid computation of the term  $\max_r \left| \sum_j \overline{u_{j+r}} v_j \right|$ . We want to show that through use of fast Fourier transform we can create a vector,  $X$ , with elements  $X(r) = \sum_j \overline{u_{j+r}} v_j$ .

**Lemma 2.** *With  $A, B \in \mathbb{C}^N$  and  $\rho = \frac{-2\pi i}{N}$ , for  $N \in \mathbb{N}$ , let*

$$X_{A,B}(r) := \frac{1}{N} \sum_{j=0}^{N-1} \left( \sum_{k=0}^{N-1} A_k e^{\rho k j} \sum_{l=0}^{N-1} B_l e^{\rho l j} \right) e^{-\rho j r}$$

for  $r \in \{0, \dots, N-1\}$ . Then

$$X_{A,B}(r) = \sum_{k+l=r} A_k B_l. \quad (\text{A.5})$$

Note that the definition of  $X_{A,B}(r)$  corresponds to the  $(r+1)$ th element of the output from running the code `ifft(fft(A).*fft(B))`, where `fft` is the fast Fourier transform and `ifft` is the inverse fast Fourier transform.

*Proof of Lemma 2.*

$$\begin{aligned} X_{A,B}(r) &:= \frac{1}{N} \sum_{j=0}^{N-1} \left( \sum_{k=0}^{N-1} A_k e^{\rho k j} \sum_{l=0}^{N-1} B_l e^{\rho l j} \right) e^{-\rho j r} \\ &= \frac{1}{N} \sum_{j=0}^{N-1} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} A_k B_l e^{\rho j(k+l-r)}. \\ &= \frac{1}{N} \left( \sum_{k+l=r} \left( \sum_{j=0}^{N-1} A_k B_l \right) + \sum_{k+l \neq r} \left( \sum_{j=0}^{N-1} A_k B_l e^{\rho j(k+l-r)} \right) \right). \end{aligned}$$

Now, if we fix  $k+l-r =: \alpha \neq 0$  and expand  $\rho$ , then we have

$$\sum_{j=0}^{N-1} A_k B_l e^{\rho j(k+l-r)} = A_k B_l \sum_{j=0}^{N-1} e^{\alpha \frac{-2\pi i}{N} j}.$$

Now note that  $\sum_{j=0}^{N-1} e^{\alpha \frac{-2\pi i}{N} j}$  is (a multiple of) the sum of all  $\frac{N}{\gcd(\alpha, N)}$ th roots of unity.

Since  $\alpha \neq 0$  we have that  $\frac{N}{\gcd(\alpha, N)} > 1$  (remember that  $\alpha$ , like all indices here, is taken mod  $N$ ). For any integer  $m > 1$  the sum of all  $m$ th roots of unity is known to be zero, hence we have

$$X_{A,B}(r) = \frac{1}{N} \left( \sum_{k+l=r} \left( \sum_{j=0}^{N-1} A_k B_l \right) \right)$$

and since  $A_k B_l$  does not depend on  $j$ .

$$X_{A,B}(r) = \sum_{k+l=r} A_k B_l. \quad (\text{A.6})$$

□

Next we wish to show that, with vector  $\overleftarrow{v}$  such that  $\overleftarrow{v}_l = v_{-l}$  (again indices are taken mod  $N$ ), we can use the computation `ifft(fft(conj(u)).*fft(overleftarrow{v}))` to give the desired vector, where `conj` represents taking the complex conjugate of elements in  $u$ .

*Proof.*

$$\begin{aligned} X_{\overline{u}, \overleftarrow{v}}(r) &= \sum_{k+l=r} \overline{u_k} \overleftarrow{v_l} \\ &= \sum_{k+l=r} \overline{u_k} v_{-l} \\ &= \sum_{k-j=r} \overline{u_k} v_j \\ &= \sum_j \overline{u_{j+r}} v_j \end{aligned}$$

□

This means we can rapidly compute a vector containing all possible values for  $|\sum_j \overline{u_{j+r}} v_j|$  and then it is simply a matter of choosing the maximum value to give us  $d(u, v)$  as in (A.4).

## A.2 Planar translation to minimise pairwise distances between two sets

Let  $a_1, \dots, a_N, b_1, \dots, b_N \in \mathbb{C}$  be points in the complex plane. Holding  $\{b_i\}$ , we want to find the planar translation of  $\{a_i\}$  that minimises the  $\ell^2$  distance between the sets, i.e. we want to find the vector  $c \in \mathbb{C}$  that minimises the expression

$$d^2 = \sum_{j=1}^N |a_j + c - b_j|^2. \quad (\text{A.7})$$

This expression can be expanded as

$$d^2 = \sum_{j=1}^N |a_j + c - b_j|^2 \quad (\text{A.8})$$

$$= \sum_{j=1}^N \left( (a_j - b_j)_R + c_R \right)^2 + \left( (a_j - b_j)_I + c_I \right)^2, \quad (\text{A.9})$$

where subscripts  $R$  and  $I$  indicate real and imaginary components respectively. Hence

$$\begin{aligned} d^2 = & Nc_R^2 + 2c_R \sum_{j=1}^N (a_j - b_j)_R + \sum_{j=1}^N (a_j - b_j)_R^2 \dots \\ & + Nc_I^2 + 2c_I \sum_{j=1}^N (a_j - b_j)_I + \sum_{j=1}^N (a_j - b_j)_I^2. \end{aligned} \quad (\text{A.10})$$

Thus,  $d^2(c)$  is a paraboloid with positive leading coefficients and we can find its minimum by looking at

$$\frac{\partial d^2}{\partial c_R} = 2Nc_R + 2 \sum_{j=1}^N (a_j - b_j)_R \quad (\text{A.11})$$

and

$$\frac{\partial d^2}{\partial c_I} = 2Nc_I + 2 \sum_{j=1}^N (a_j - b_j)_I. \quad (\text{A.12})$$

Setting both partial derivatives to zero yields the solution,

$$c_R + ic_I = \frac{1}{N} \sum_{j=1}^N \left( (b_j - a_j)_R + i(b_j - a_j)_I \right) \quad (\text{A.13})$$

$$c = \frac{1}{N} \sum_{j=1}^N (b_j - a_j) = \frac{1}{N} \sum_{j=1}^N (b_j) - \frac{1}{N} \sum_{j=1}^N (a_j). \quad (\text{A.14})$$

Hence, the translation that minimises the distance between pairs of points is that which aligns the means of the two sets.



# Bibliography

- M Abercrombie, J E Heaysman, and S M Pegrum. The locomotion of fibroblasts in culture. I. Movements of the leading edge. *Experimental cell research*, 59(3): 393–8, March 1970. ISSN 0014-4827.
- Maria Teresa Abreu-Blanco, James J. Watts, Jeffrey M. Verboon, and Susan M. Parkhurst. Cytoskeleton responses in wound repair. *Cellular and Molecular Life Sciences*, 69(15):2469–2483, 2012. ISSN 1420-682X. doi: 10.1007/s00018-012-0928-2.
- Natalie Andrew and Robert H Insall. Chemotaxis in shallow gradients is mediated independently of PtdIns 3-kinase by biased choices between random protrusions. *Nature cell biology*, 9(2):193–200, March 2007. ISSN 1465-7392. doi: 10.1038/ncb1536.
- Chris Bakal, John Aach, George Church, and Norbert Perrimon. Quantitative morphological signatures define local signaling networks regulating cell morphology. *Science (New York, N.Y.)*, 316(5832):1753–6, June 2007. ISSN 1095-9203. doi: 10.1126/science.1140324.
- LE Baum and T Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *The annals of mathematical statistics*, pages 1554–1563, 1966.
- S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, April 2002. ISSN 01628828. doi: 10.1109/34.993558.
- Guillaume Charras and Ewa Paluch. Blebs lead the way: how to migrate without lamellipodia. *Nature reviews. Molecular cell biology*, 9(9):730–6, September 2008. ISSN 1471-0080. doi: 10.1038/nrm2453.
- Guillaume T Charras, Justin C Yarrow, Mike A Horton, L Mahadevan, and T J

- Mitchison. Non-equilibration of hydrostatic pressure in blebbing cells. *Nature*, 435(7040):365–9, May 2005. ISSN 1476-4687. doi: 10.1038/nature03550.
- R Coifman and S Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, July 2006a. ISSN 10635203. doi: 10.1016/j.acha.2006.04.006.
- Ronald R. Coifman and Stéphane Lafon. Geometric harmonics: A novel tool for multiscale out-of-sample extension of empirical functions. *Applied and Computational Harmonic Analysis*, 21(1):31–52, July 2006b. ISSN 10635203. doi: 10.1016/j.acha.2005.07.005.
- M S Cooper and M Schliwa. Motility of cultured fish epidermal cells in the presence and absence of direct current electric fields. *The Journal of cell biology*, 102(4): 1384–99, April 1986. ISSN 0021-9525.
- T F Cootes, C J Taylor, D H Cooper, and J Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995. ISSN 10773142. doi: 10.1006/cviu.1995.1004.
- Louise P Cramer. Forming the cell rear first: breaking cell symmetry to trigger directed cell migration. *Nature cell biology*, 12(7):628–32, July 2010. ISSN 1476-4679. doi: 10.1038/ncb0710-628.
- Gaudenz Danuser, Jun Allard, and Alex Mogilner. Mathematical modeling of eukaryotic cell migration: insights beyond experiments. *Annual review of cell and developmental biology*, 29:501–28, January 2013. ISSN 1530-8995. doi: 10.1146/annurev-cellbio-101512-122308.
- Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *Science (New York, N.Y.)*, 315(5814):972–6, February 2007. ISSN 1095-9203. doi: 10.1126/science.1136800.
- Peter Friedl and Katarina Wolf. Tumour-cell invasion and migration: diversity and escape mechanisms. *Nature reviews. Cancer*, 3(5):362–74, May 2003. ISSN 1474-175X. doi: 10.1038/nrc1075.
- HB Goodrich. Cell behavior in Tissue Cultures. *Biological Bulletin*, pages 252–262, 1924.
- Kristen Grauman and Trevor Darrell. Fast Contour Matching Using Approximate Earth Movers Distance. (June), 2004.

- Douglas Hanahan, Robert A Weinberg, and San Francisco. The Hallmarks of Cancer Review University of California at San Francisco. 100:57–70, 2000.
- Rick Horwitz and Donna Webb. Cell migration. *Current biology : CB*, 13(19): R756–9, September 2003. ISSN 0960-9822.
- Andrew R Houk, Alexandra Jilkine, Cecile O Mejean, Rostislav Boltyanskiy, Eric R Dufresne, Sigurd B Angenent, Steven J Altschuler, Lani F Wu, and Orion D Weiner. Membrane tension maintains cell polarity by confining signals to the leading edge during neutrophil migration. *Cell*, 148(1-2):175–88, January 2012. ISSN 1097-4172. doi: 10.1016/j.cell.2011.10.050.
- F. Huber, J. Schnauß, S. Röncke, P. Rauch, K. Müller, C. Fütterer, and J. Käs. Emergent complexity of the cytoskeleton: from single filaments to tissue. *Advances in Physics*, 62(1):1–112, February 2013. ISSN 0001-8732. doi: 10.1080/00018732.2013.771509.
- Samuel D R Jefferyes, David B A Epstein, Anne Straube, and Nasir M Rajpoot. A novel framework for exploratory analysis of highly variable morphology of migrating epithelial cells. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2013:3463–6, January 2013. ISSN 1557-170X. doi: 10.1109/EMBC.2013.6610287.
- Shantanu H Joshi, Eric Klassen, Anuj Srivastava, and Ian Jermyn. Removing Shape-Preserving Transformations in Square-Root Elastic (SRE) Framework for Shape Analysis of Curves. *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 4679:387–398, January 2007. ISSN 1063-6919. doi: 10.1007/978-3-540-74198-5\_30.
- R Keller. Cell migration during gastrulation. *Current opinion in cell biology*, 2005.
- Kinneret Keren, Zachary Pincus, Greg M Allen, Erin L Barnhart, Gerard Marriott, Alex Mogilner, and Julie A Theriot. Mechanism of shape determination in motile cells. *Nature*, 453(7194):475–480, 2008.
- Kinneret Keren, Patricia T Yam, Anika Kinkhabwala, Alex Mogilner, and Julie A Theriot. Intracellular fluid flow in rapidly moving cells. *Nature cell biology*, 11(10):1219–24, October 2009. ISSN 1476-4679. doi: 10.1038/ncb1965.

- Adnan Mujahid Khan, Hesham Eldaly, and Nasir M Rajpoot. A gamma-gaussian mixture model for detection of mitotic cells in breast cancer histopathology images. *Journal of pathology informatics*, 4:11, January 2013. ISSN 2229-5089. doi: 10.4103/2153-3539.112696.
- Olcay Kursun. Spectral Clustering with Reverse Soft K-Nearest Neighbor Density Estimation. *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2010. doi: 10.1109/IJCNN.2010.5596620.
- Stéphane Lafon and Ann B Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1393–1403, 2006.
- Longin Jan Latecki. MPEG 7 Shape Matching. <http://www.dabi.temple.edu/~shape/MPEG7/dataset.html>.
- D A Lauffenburger and A F Horwitz. Cell migration: a physically integrated molecular process. *Cell*, 84(3):359–69, February 1996. ISSN 0092-8674.
- J Lee, A Ishihara, JA Theriot, and K Jacobson. Principles of locomotion for simple-shaped cells. *Letters to nature*, 1993.
- Gary P Liney, Muthyala Sreenivas, Peter Gibbs, Roberto Garcia-Alvarez, and Lindsay W Turnbull. Breast lesion analysis of shape technique: semiautomated vs. manual morphological description. *Journal of magnetic resonance imaging : JMRI*, 23(4):493–8, April 2006. ISSN 1053-1807. doi: 10.1002/jmri.20541.
- Federica M Marelli-Berg, Hongmei Fu, Fabrizio Vianello, Koji Tokoyoda, and Alf Hamann. Memory T-cell trafficking: new directions for busy commuters. *Immunology*, 130(2):158–65, June 2010. ISSN 1365-2567. doi: 10.1111/j.1365-2567.2010.03278.x.
- Sophie G Martin and Martine Berthelot-Grosjean. Polar gradients of the DYRK-family kinase Pom1 couple cell length with the cell cycle. *Nature*, 459(7248):852–6, June 2009. ISSN 1476-4687. doi: 10.1038/nature08054.
- Washington Mio, Anuj Srivastava, and Shantanu Joshi. On Shape of Plane Elastic Curves. *International Journal of Computer Vision*, 73(3):307–324, September 2007. ISSN 0920-5691. doi: 10.1007/s11263-006-9968-0.
- T J Mitchison and L P Cramer. Actin-based cell motility and cell locomotion. *Cell*, 84(3):371–9, February 1996. ISSN 0092-8674.

- Alex Mogilner and Leah Edelstein-Keshet. Regulation of Actin Dynamics in Rapidly Moving Cells: A Quantitative Analysis. *Biophysical Journal*, 83(September), 2002.
- Alex Mogilner and Jie Zhu. Cell polarity: tension quenches the rear. *Current biology : CB*, 22(2):R48–51, January 2012. ISSN 1879-0445. doi: 10.1016/j.cub.2011.12.013.
- James B Moseley, Adeline Mayeux, Anne Paoletti, and Paul Nurse. A spatial gradient coordinates cell size and mitotic entry in fission yeast. *Nature*, 459(7248): 857–60, June 2009. ISSN 1476-4687. doi: 10.1038/nature08074.
- Kevin Murphy and Matt Dunham. Probalistic modeling toolkit. URL <https://github.com/probml/pmtk3>.
- K Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical ...*, pages 559–572, 1901.
- Thomas D Pollard and Gary G Borisy. Cellular motility driven by assembly and disassembly of actin filaments. *Cell*, 112(4):453–65, February 2003. ISSN 0092-8674.
- Neta Rabin and RR Coifman. Heterogeneous Datasets Representation and Learning using Diffusion Maps and Laplacian Pyramids. *SDM*, pages 575–585, 2012.
- Nasir Rajpoot and Muhammad Arif. Unsupervised Shape Clustering using Diffusion Maps. *Applied Sciences*, 2008(5):1–16, 2008.
- Rangaraj M Rangayyan and Thanh M Nguyen. Fractal analysis of contours of breast masses in mammograms. *Journal of digital imaging*, 20(3):223–37, September 2007. ISSN 0897-1889. doi: 10.1007/s10278-006-0860-9.
- Raphaela Rid, Natalia Schiefermeier, Ilya Grigoriev, J Victor Small, and Irina Kaverina. The last but not the least: the origin and significance of trailing adhesions in fibroblastic cells. *Cell motility and the cytoskeleton*, 61(3):161–71, July 2005. ISSN 0886-1544. doi: 10.1002/cm.20076.
- Anne J Ridley, Martin A Schwartz, Keith Burridge, Richard A Firtel, Mark H Ginsberg, Gary Borisy, J Thomas Parsons, and Alan Rick Horwitz. Cell migration: integrating signals from front to back. *Science (New York, N.Y.)*, 302(5651):1704–9, December 2003. ISSN 1095-9203. doi: 10.1126/science.1092053.

- Julia Riedl, Alvaro H Crevenna, Kai Kessenbrock, Jerry Haochen Yu, Dorothee Neukirchen, Michal Bista, Frank Bradke, Dieter Jenne, Tad A Holak, Zena Werb, Michael Sixt, and Roland Wedlich-Soldner. Lifeact: a versatile marker to visualize F-actin. *Nat Meth*, 5(7):605–607, July 2008. ISSN 1548-7091.
- Evanthia T Roussos, John S Condeelis, and Antonia Patsialou. Chemotaxis in cancer. *Nature reviews. Cancer*, 11(8):573–87, August 2011. ISSN 1474-1768. doi: 10.1038/nrc3078.
- Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. A Metric for Distributions with Applications to Image Databases. 1998.
- Alon Schclar. A diffusion framework for dimensionality reduction. In Oded Maimon and Lior Rokach, editors, *Soft Computing for Knowledge Discovery and Data Mining*, pages 315–325. Springer US, 2008. ISBN 978-0-387-69934-9.
- Rachel Sparks and Anant Madabhushi. Explicit shape descriptors: novel morphologic features for histopathology classification. *Medical image analysis*, 17(8):997–1009, December 2013. ISSN 1361-8423. doi: 10.1016/j.media.2013.06.002.
- Anuj Srivastava, Eric Klassen, Shantanu H Joshi, and Ian H Jermyn. Shape Analysis of Elastic Curves in Euclidean Spaces. *IEEE transactions on pattern analysis and machine intelligence*, 33(7):1415–1428, September 2011. ISSN 1939-3539. doi: 10.1109/TPAMI.2010.184.
- Olaf Strauss. The retinal pigment epithelium in visual function. *Physiological reviews*, 85(3):845–81, July 2005. ISSN 0031-9333. doi: 10.1152/physrev.00021.2004.
- Katsiaryna Tarbashevich and Erez Raz. The nuts and bolts of germ-cell migration. *Current opinion in cell biology*, 22(6):715–21, December 2010. ISSN 1879-0410. doi: 10.1016/j.ceb.2010.09.005.
- Ulrike Theisen, Ekkehard Straube, and Anne Straube. Directional Persistence of Migrating Cells Requires Kif1C-Mediated Stabilization of Trailing Adhesions. *Developmental cell*, pages 1153–1166, 2012.
- Eric Thevenneau and Roberto Mayor. Collective cell migration of the cephalic neural crest: the art of integrating information. *Genesis (New York, N.Y. : 2000)*, 49(4):164–76, April 2011. ISSN 1526-968X. doi: 10.1002/dvg.20700.

- Luke Tweedy, Börn Meier, Jürgen Stephan, Doris Heinrich, and Robert G Endres. Distinct cell shapes determine accurate chemotaxis. *Scientific reports*, 3:2606, January 2013. ISSN 2045-2322. doi: 10.1038/srep02606.
- AJ Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, pages 260–269, 1967.
- Hong-Rui Wang, Yue Zhang, Barish Ozdamar, Abiodun A Ogunjimi, Evguenia Alexandrova, Gerald H Thomsen, and Jeffrey L Wrana. Regulation of cell polarity and protrusion formation by targeting RhoA for degradation. *Science (New York, N.Y.)*, 302(5651):1775–9, December 2003. ISSN 1095-9203. doi: 10.1126/science.1090772.
- Weigang Wang, Sumanta Goswami, Erik Sahai, Jeffrey B Wyckoff, Jeffrey E Segall, and John S Condeelis. Tumor cells caught in the act of invading: their strategy for enhanced cell motility, 2005. ISSN 09628924.
- Joe H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58:236244, 1963.
- Gregory F Weber, Maureen A Bjerke, and Douglas W DeSimone. A mechanoresponsive cadherin-keratin complex directs polarized protrusive behavior and collective cell migration. *Developmental cell*, 22(1):104–15, January 2012. ISSN 1878-1551. doi: 10.1016/j.devcel.2011.10.013.
- Orion D Weiner. Regulation of cell polarity during eukaryotic chemotaxis: the chemotactic compass. *Current Opinion in Cell Biology*, 14(2):196–202, April 2002. ISSN 09550674. doi: 10.1016/S0955-0674(02)00310-1.
- David Wishart. ClustanGraphics3 Interactive Graphics for Cluster Analysis Seriation Algorithm, 1999.
- Patricia T Yam, Cyrus A Wilson, Lin Ji, Benedict Hebert, Erin L Barnhart, Natalie A Dye, Paul W Wiseman, Gaudenz Danuser, and Julie A Theriot. Actin-myosin network reorganization breaks symmetry at the cell rear to spontaneously initiate polarized cell motility. *The Journal of cell biology*, 178(7):1207–21, September 2007. ISSN 0021-9525. doi: 10.1083/jcb.200706012.
- Jing Yang and Robert A Weinberg. Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis. *Developmental cell*, 14(6):818–29, June 2008. ISSN 1878-1551. doi: 10.1016/j.devcel.2008.05.009.

- Wei Yang, Su Zhang, Yazhu Chen, Wenying Li, and Yaqing Chen. Shape symmetry analysis of breast tumors on ultrasound images. *Computers in biology and medicine*, 39(3):231–8, March 2009. ISSN 1879-0534. doi: 10.1016/j.combiomed.2008.12.007.
- Laurent Younes. Computable elastic distance between shapes. *SIAM Journal of Applied Mathematics.*, 58:565586, 1998.
- L Zelnik-Manor and P Perona. Self-tuning spectral clustering. *NIPS Proceedings*, 2:1601–1608, 2005.